Quality & Quantity (2024) 58:603–625
https://doi.org/10.1007/s11135-023-01654-3

Using web archives for an explorative study of the web presence of German parties during the European election 2019

Florence Ertel[1] · Simon Donig[2] · Markus Eckl[3] · Sebastian Gassner[2] · Daniel Göler[1] · Malte Rehbein[2]

Accepted: 14 March 2023 / Published online: 12 April 2023

Abstract
In the digital age, political science is faced with a shift of election campaigns and political discourse to digital or virtual arenas. Because the internet is a highly volatile medium and online content can become inaccessible after the campaign season, new challenges for research arise as well as the need for the preservation of online content. Moreover, the sheer volume of data researchers have to deal with has reached levels where traditional methods are being highly challenged. This paper puts forth a web harvesting workflow with a strong focus on granular extraction of unstructured information (publication dates) for automated analysis. As our approach is methodological, we would like to point out the benefits that researches in political science may draw from adapting our methodology. We demonstrate this by analysing an event-based web crawl of German parties participating in the election campaign for the European Parliamentary Election in 2019. We employ distant reading methods to generate topic models, which are subsequently evaluated by hermeneutic analysis of a subset of the data.

Keywords Web archives · European parliamentary election · Second order elections · Distant reading · Topic modelling

1 Introduction

In the digital age, political science is faced with a shift of election campaigns and political discourses to digital or virtual arenas. This leads to different methodological challenges. Firstly, the internet is a highly volatile medium and online content can become inaccessible after the campaign season. This creates a need for a form preservation of online content. This becomes especially pertinent with regard to academic standards and the need for research

✉ Florence Ertel
florence.ertel@uni-passau.de

[1] Jean-Monnet-Chair of European Politics, University of Passau, 94030 Passau, Germany

[2] Chair of Digital Humanities, University of Passau, 94030 Passau, Germany

[3] Department of Social Work, University of Applied Sciences Fulda, 36037 Fulda, Germany

to be both replicable and erifiably reliable. Secondly, the volume of online data researchers have to content with has reached levels which challenge the efficacy of traditional data collection and processing methods. In this paper, we address these challenges and follow a mainly methodological research interest with an empirical application to the European Parliament elections in 2019. While web scraping techniques are already widely used in the political sciences, we focus on an alternative and explorative approach based on web archives[1] as a standardised format for archiving and working with crawl data.[2] We seek to showcase the benefits that disciplines such as political science can derive from both adapting web harvesting with standardised archival formats, and the granular extraction of unstructured temporal data from web archives. These methods facilitate the structuring of archival content for further analysis with digital instruments in mixed methods research designs. Our approach makes use of a combination of data collection, data storage, data processing (archiving) and analysis (through text mining methods). In order to demonstrate the efficacy of our methodological approach, we will provide a case study on the election campaign for the European Parliament in 2019. Although our main concern here is methodological, we recognise that a substantive research question is indispensable for the conception of the methodologically oriented approach. Against the backdrop of research on the second-order character of European elections based on Reif and Schmitt (1980), we want to use an exploratory approach to detect which issues frequently emerged in the 2019 European election campaign.

We concentrate on Web archiving as a data collection and data storage method, because it offers standardised data formats that can easily be passed on and reused in multiple ways, from reliving the browsing experience of the past web to extracting and analysing linkages between websites or extracting and enriching website content. In contrast, Web scraping can be used as a method to create structured data, but such data is usually not standardised and hard to reuse outside of the original context and purpose. The underlying technique of web crawling and content acquisition can be fully- or semi-automated, or entirely human-driven. While web archiving is a technology that can be used by anyone, traditional memory conservation and consolidation institutions such as libraries and archives have stepped in to preserve parts of the web. Remarkably, though, the Internet Archive, a non-profit organisation, has grown into one of the largest archives of web-content (514 billion web pages at time of writing) since 1996. It provides one of the most popular interfaces to interact with the archived web, the so-called Wayback Machine (https://web.archive.org/). In the academic sphere, historians have been among the keenest adopters of web archiving as a research technique. Consequently, there is an extensive amount of literature on using web archives as means of writing web history (e.g. Brügger and Milligan 2018).

However, in our study we do not focus on existing web archives whose composition cannot be influenced by the researcher, but carry out the web archiving process ourselves. The data corpus was deliberately chosen to be limited in order to exemplify the entire process—from data collection, storage and extraction to the analysis using text mining methods.

[1] Web archives provide a standardised format for the preservation of snapshots of online content (Brügger and Milligan 2019).

[2] The term web scraping describes the automatic extraction of content from a website. For this purpose, instruments such as the wget tool are used. The data formats generated by web scraping can be quite granular and even reusable, but they are not standardised and usually insufficiently enriched with metadata about the actual scraping process. The quality of the data depends to a large extent on the individual researcher's ambition to produce reusable data and is therefore inconsistent.

Web archiving makes use of various methods to crawl web content. The crawling process is documented in a standardised archive format. At the same time, the archived content of a web archive is usually more comprehensive and extensive than that generated by a scraping process.

To that end, we analyse the web pages of eleven German parties running for the European elections in 2019. The dataset consists of web archives derived from the aforementioned websites. These archives have been built through event crawls with web harvesting tools, based on an actor-centred seed list. In this paper, we test the combined use of web crawling, granular data extraction, especially of unstructured temporal text data, and quantitative text analysis techniques. The latter will be focussed on topic modelling, by studying which issues are reflected on the web presences of selected German political parties during the European election campaign 2019.

2 Case study: European Parliament elections 2019

In this case study, our research interest is mainly methodological. We illustrate our approach focussing on the internet campaigns of German political parties in the run up for the European elections 2019 by providing a combination of data collection, data storage, data processing (archiving) and analysis. We therefore analyse the data collected by web harvesting with standardised archival formats with distant reading methods, namely topic modelling, to detect which issues were of importance in the 2019 European election campaign. We thus utilise a primarily explorative approach to test which topics commonly feature, or even dominate, the communication on party websites and venture to place these in the larger context of research into European parliament elections. Referring to the first- and second-order election model (Reif and Schmitt 1980; Irwin 1995; Reif 1997; Träger 2014), which can be considered one of the most used concepts for analysing European Parliament (EP) elections (Holtz-Bacha 2020; Schmitt and Toygür 2016; Braun and Schmitt 2020; Haßler et al. 2019a, b; Haußner and Kaeding 2019; Ehin and Talving 2021; Ehin and Talving2020, Holtz-Bacha 2020;Schmitt et al. 2020), we seek to draw conclusions about the importance parties attribute to the European elections through their website communication and use the theory as an extended interpretative framework for our analytical results. By doing so we are referring to the second-order-election research branch which is focusing on the issues of the campaigns and its media coverage (Galpin and Trenz 2019; Schulte-Cloos 2018; Braun and Schwarzbözl 2019) and which is also labelled as "second-rate election campaigning" (de Vreese 2009). Particular importance is payed to the question, whether the campaigns are "dominated by domestic issues [or] issues on the EU agenda" (Hix and Marsh 2011: 4). By answering this question, we would like to contribute to the broader debate about, whether the second order election model still possesses explanatory power (Schmitt and Toygür 2016: 176; Ehin and Talving 2021) or whether "the traditional distinction between first- and second-order election is insufficient to grasp the public sphere dynamics of politicised EP elections" (Galpin and Trenz 2019: 1, see also Galpin and Trenz 2019; Gattermann, de Vreese and van der Brug 2021). By analysing internet-based data, we want to connect the discussion on the opportunities of the so-called digital revolution in social science and the "avalanche of data related to politics" (Wilkerson and Casas 2017: 530) that the world wide web offers (Kaiser 2014; Alvares 2016; Wilkerson and Casas 2017). Furthermore, this analysis addresses the methodological challenges resulting from this revolution (Karpf 2012; Wilkerson and Casas 2017), which occur on two different levels. On the one hand, challenges arise from the collection and storage or archiving of collected data in ways, which are compliant with academic standards (Göler and Reiter 2019). Dealing with the challenges for data acquisition and archiving of internet-based data, which are characterised by a high degree of "fluidity", is a key issue in academic research using the internet as a data

source. Web archiving is central to ensure replicability and reliability when working with this kind of data, as collecting, storing, and archiving web data cannot be covered by traditional methods in a reliable manner. While web scraping can be used as a method to create structured data, such data is usually not standardised and hard to reuse outside of a original research context and purpose. In contrast, web archiving offers standardised data formats that can easily be passed on and reused in multiple ways, from reliving the browsing experience of the past web to extracting and analysing linkages between websites or extracting and enriching website content. By archiving websites, researchers can both document and analyse the changes in communication on the internet. Web archiving thus enables researchers "to document our findings when we study today's web, since in practice most web studies preserve the web in order to have a stable object to study and refer to when the analysis is to be documented (except for studies of the live web)" (Brügger 2011: 24). Due to the shift of social and political communication processes to the internet, web archiving as a research process and web archives as data bases for analyses are not only indispensable for web historians, such as Brügger, but also for scholars in other research fields.

On the other hand, challenges arise from handling this large amount of data. As Grimmer and Steward (2013) point out, automated text analysis—which is frequently used for exploring large amounts of data—requires "careful thought and reasoning" (Grimmer and Steward 2013: 295). Internet-based text as data offers a wide range of opportunities for research and "has produced important advances in research methods" (Wilkerson and Casas 2017: 540). For example, in European integration studies, research on political and social debates cannot be conducted without incorporating both the party's and politician's websites and social media platforms. Since the so-called migration crisis, more weight is placed to questions of integration research, especially from a post-functionalist perspective (Hooghe and Marks 2019: 1122). European integration research that analyses these developments without considering the importance of web-based data would ignore a part of today's social reality. Additionally, we can see a rising success of populist and EU-sceptic parties across Europe. Since these parties are often making extensive use of social media in their communication strategies (Schaub and Morisi 2019; Grill 2016) and their electorate commonly uses the internet as their primary source of information (Maggini 2014: 57), research focussing on these parties can hardly be conducted without reference to internet-based data. Thus, the importance of social media platforms for campaigning communication is growing steadily and the interest in analysing this kind of data is obvious (Bentivegna and Marchetti 2014; Barberà et al. 2019; Marchal et al 2019; Valentini 2019; Haußner and Klika 2019; Pfaffenberger and Heinrich 2020; Haßler et al. 2019a, b). Nevertheless, the analysis of party websites also remains highly relevant when investigating the online communication of political actors (Rußmann 2016: 56). As Rußmann notes, communication on websites is about presenting an overall image of the party or a specific politician. Website analyses therefore show the extent to which political actors use websites to spread information. Compared to social media platforms like Twitter, Instagram or Facebook, websites offer nearly unlimited space for information (Rußmann 2016: 56). Furthermore, party websites serve as useful examples of dealing with standard websites and developing adequate archiving and data extraction techniques, as they differ for example in their structure and in the way, they incorporate external content.

In our analysis we have therefore deliberately limited ourselves to party websites in order to map the entire process of data collection—extraction, storage and analysis. However, this choice of instrument has implications for the use of embedded social media, which is usually not captured by current crawling techniques. Based on dynamic web technologies such as java script, social media content is usually loaded when the page is

rendered in the browser and thus—by its embedded nature—harder to capture with standard crawling techniques. Moreover, social media content is usually of a limited length in characters, making it difficult to compare it with full length webpages by means of statistical methods such as topic modelling. For the crawling of the party websites, we were supported by the Bavarian State Library, which crawled the web pages during the election campaign according to its scientific standards, our predefined selection criteria and technical standards that guaranteed the usability for our methodological toolkit.

3 Methodological considerations and applied methodological approach

Studying the content of web archives poses epistemological as well as methodological challenges. To analyse information contained in web archives comprehensively, advanced methods of information retrieval and data analysis are required. Because there is no established best practice, we see our approach as presented and discussed here as a contribution to establishing new techniques to analyse web data.

Firstly, the study of archived web content requires a reliable and reproducible way of extracting data from a web archive and enrich it with metadata from the archive. Here we went beyond the state of the art by developing a specific piece of software called *warc-2corpus (W2C)* that facilitates the finely grained extraction of data from Web ARChive (WARC) files. Secondly, the extracted and enriched data needs to be structured and analysed in a way that will make it beyond the close reading of singular web pages. To this end we opted for a distant reading approach. In this regard, we rely on the technique of topic modelling, which is already widely used in disciplines like communication studies or literary studies. Particularly, we apply a highly promising distant reading method called Structural Topic Modelling (STM), allowing us to investigate topics in relation to variables external to the text as such, and derived from the web archive, such as time or party. Finally, distant reading techniques like STM allow us to explore and analyse large amounts of text in contrast to traditional close reading methods, where the researcher actually reads the text. As the results obtained from the STM need to be contextualised, we perform this intellectual examination with a random reading of selected texts from the crawled websites.

3.1 Distant reading

In order to access texts and their contents, political science generally uses various qualitative methods, ranging from content analysis and hermeneutic understanding of meaning to grounded theory. With the help of these methods, important phenomena can be discovered and analysed. However, qualitative text analysis reaches its limits when it is necessary to analyse larger text corpora efficiently (Jannidis et al. 2017). Due to the enormous increase in digitised or digitally born texts, i.e. texts that are already digital at the time of their creation, the textual datasets to be analysed also grow enormously (Brynjolfsson and McAfee 2014), as can be observed in the context of web hierarchies. In order to examine this flood of text beyond a random or selective basis, there is a growing need to tackle the challenges in using quantitative methods of text analysis.

In our analysis, we apply a fundamentally different approach to the use of quantitative methods developed by Moretti (2013). He established the notion of distant (as opposed to close) reading as a method for dealing with the "great unread" of non-canonical literature

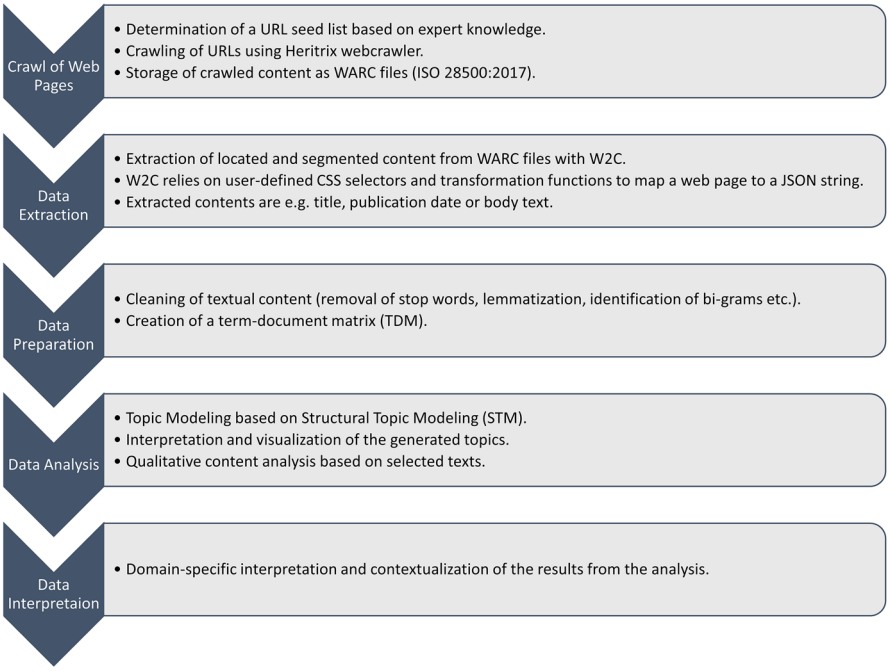

Fig. 1 Workflow

by digital means (Moretti 2013). It seems important to note that distant reading is not supposed to replace close reading, but rather to expand the study of text with a new dimension. Thus, distance can serve "as a condition of knowledge that allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems" (Moretti 2013: 48–49). It can be used to deal with new and different questions than those based on close reading. Hermeneutic contexts of meaning or the reconstruction of specific individual behaviour patterns cannot be determined to the extent that is known from qualitative research, but texts can be structured with the help of algorithms and can thus be converted for mathematical operations (Stulpe and Lemke 2016). Texts can then be arranged, edited, sorted or searched for content, i.e. structures and recurring patterns can be researched (Hockey 2000; Heyer et al. 2006: 3). In this case study, we use the method of topic modelling, more precisely Latent Dirichlet Allocation (LDA), which can be understood as a distant reading method (Stulpe and Lemke 2016). Before we go into more detail, we will first explain how the data extraction from the web archive was done and how a corpus was built.

3.2 Event-crawling the European Elections of 2019, WARC and warc2corpus

When we go into our workflow in more detail below, we would like to highlight five stages. These are the crawl of web pages, data extraction, data preparation, data analysis and data interpretation. Figure 1 shows our workflow and summarizes the essential aspects of each step.

We built a corpus for the European Election of 2019 by crawling websites of political parties campaigning in this election, crawling each website roughly on a daily basis between 22 March 2019 and 1 July 2019. This strategy of content acquisition is called an event crawl because it is directed at a short period of interest rather than long term archiving of a target (Brügger 2018: 20). For that purpose, we initialised the crawl by providing a set of URLs that serve as an entry-point for the crawler, a seed list. Starting with the seed URL, for instance the website of the German party Bündnis 90/Die Grünen (www.gruene.de), a crawler will then randomly follow the URLs contained in that page. The crawl can be limited to a particular domain and will stop after a predefined number of links have been visited.

We chose party websites as objects of investigation because, as Rußmann notes, they are part of the standard repertoire of every party's campaign tools (Rußmann 2016: 55). Although the analysis of social media channels has increasingly become the focus of research for a good decade, the analysis of actors' websites remains highly relevant in election campaigns. The 2009 federal and 2009 European election campaigns saw the first intensive use of social media and researchers increasingly focus on these channels when analysing web campaigns. Nevertheless, party and politician websites provide an adequate data basis to study election campaign communication. While communication on social media channels must follow certain rules, such as a limit of 280 Unicode characters for tweets on Twitter or a photo or short video combined with appropriate hashtags on Instagram, there is almost unlimited space for information on a classic website. By communicating on websites, actors can present an overall image of their party or a specific politician. Website analyses accordingly show the extent to which political actors use websites in online election campaigns (Lilleker and Jackson 2011; Rußmann 2016).

Regarding the investigation period, the last four weeks before the election day are considered as critical phase in political communication research (Woyke 2013: 133). For this reason, we focus on the period from 23 April 2019 until the German election day 26 May 2019 (the election period in Europe was from 23 to 26 May 2019) and the two following weeks for the event crawl. By using this period, we were sure to cover the most intensive period of the campaign as well as the first reactions of the parties to the outcome of the election. The most popular standard for archiving web content are so-called Web ARChive (WARC) files. Each crawl of a single target, i.e. a political party website, results in a single WARC file generated for that particular crawl and this specific target. The file format offers a convention for accommodating a broad variety of resources, such as text, images, video or other formats. The WARC format stores "webcrawls" as sequences of content-blocks harvested from the World Wide Web. Each capture in a WARC file is preceded by a header recording meta-data on the harvested content, followed by the retrieval protocol response messages and the content itself. WARC can furthermore accommodate secondary content, such as assigned metadata, abbreviated duplicate detection events, later-date transformations, and segmentation of large resources (ISO 2017). After obtaining a set of WARCs, we continued by identifying and extracting a relevant subset of the web pages contained in each web archive. For instance, we disregarded all web pages containing information such as contact data or legal notices already in this initial phase. While a web page may look uniform when rendered in a browser, it is in fact only moderately standardised. This standardization is chiefly focused on the appearance of web content and to a much lesser degree on the underlying structure of each web page. In order to process information with a computer, this loosely and heterogeneously structured data (so called un- or semi-structured information) needs to be turned into structured information making the computer

aware what the exact relation between each piece of information and a particular electronic resource within the site it belongs to is.

Our interest was focused on electronic resources that we label as an document. Documents is a term we borrow from the topic modelling community to describe each textual surrogate associated with a particular date from the web archive. A document is a non-distinct smallest unit of analysis in topic modelling supposed to be composed of one or multiple topics. We define a document as having at minimum a title, a date of publication and a body of text; additionally, an document may comprise author(s), subtitle, teaser text, etc. As all this information may be found in no fixed order inside a web page, a transformation into the structured format defined as an document is required. This poses several challenges. For instance, as Kohlschütter et al. (2010) note, in addition to the content displayed to the user, web pages consist of navigational elements, templates, and advertisements. This boilerplate text typically is not related to the main content, may deteriorate search precision, and thus needs to be detected properly and be removed. Since our aim is to analyse the diffusion of topics of political discussion in the course of time, we also have to assign the date of publication to each document. As Stevenson and Ben-David put it:

> A [...] challenge in applying historical network analyses of the web relates to the difficulty in determining websites' specific timestamps. The live web hosts various websites that were published or recently updated at different points in time, and the exact time of publication or update can be estimated through various methods, such as the timestamps embedded in a website's code, or a response to a HTTP server request. (Stevenson and Ben-David 2018: 132)

To address these challenges, we developed an open-source tool named warc2corpus (W2C) (https://github.com/sepastian/warc2corpus). W2C is a Python library which can be used stand-alone or from within the Archives Unleashed Toolkit (AUT) (Archives Unleashed Project 2019).[3] W2C relies on user-defined CSS selectors and transformation functions to map a web page to a JSON string representing an document. Unlike other, heuristic approaches for boilerplate removal and determining the age of a web page, W2C extracts precise, structured information at the cost of additional configuration. Writing configuration for each version of a website is demanding, but it is required only initially; whenever the layout of a web page changes significantly, the configuration may require an update. Once in place, the user-defined configuration can be used to process any number of web pages at no additional cost. The configuration consists mainly of so-called extractors and is written in Python 3. An extractor specifies the name a piece of information to obtain as well as its location inside a web page; optionally, a mapping function applied on the data extracted may be specified. This way, the extractor has been used to obtain an arbitrarily formatted date from a known location inside the web page and generate a string containing a date conforming to ISO 8601. Extractors can be defined for all relevant pieces of information. In our analysis, we extract and use title, date of release and body text for each document. After extracting all web pages with W2C, we need to deduplicate the resulting documents because the same URL may have been visited several times during subsequent crawls, a web page may have been accessible through different URLs, or the same content may have been provided in different versions for both desktop browsers and mobile devices. Finally, the documents extracted can be used as input data for topic modelling.

[3] The AUT is an open-source platform delivering tools for analysing and processing large scale webcrawls, building on the Apache Spark infrastructure.

3.3 Topic modelling: LDA and STM

In recent years, topic modelling as a method of quantitative analysis of texts has been discussed frequently, especially in the Digital Humanities (McFarland et al. 2013) and the computational social sciences (Rodriguez and Storer 2020). Topic modelling is generally considered a generative model, which reproduces the construction of a text with the aid of probabilistic methods (Steyvers and Griffiths 2007). The aim of such methods is to identify latent semantic structures in the texts and thereby extract topics. They are particularly suitable for the analysis of very large amounts of text data for which the contents are largely unknown or for which the data are characterised by unstructured and missing metadata. Besides the Latent Semantic Analysis (LSA) by Deerwester et al. (1990), the Latent Dirichlet Allocation by Blei et al. (2003) is probably the best-known method of topic modelling. With the help of LDA, topics of the European election campaign are to be explored, which can then be examined to see whether they address national or European issues.

LDA is characterized by the fact that scholars do not have to subjectively specify topics, i.e. in the form of keywords whose frequency is determined in a text. With LDA, topics can be found based on common words in the documents using a complex probabilistic model (Blei et al. 2003). The starting point of LDA is the determination of two probability distributions. Thus, there is a probability distribution that in the individual texts (in the following: documents), certain topics are included. Two documents are defined as similar if both contain the same words. Each time a new document is added, the algorithm checks how closely it matches the other documents in terms of wording. The second probability distribution contains the probability that a topic consists of certain words. The basic assumption behind these two probability distributions is that a document has only a limited number of topics and that each topic can only be represented by certain words from the documents.

For LDA, the common occurrence of words, also called co-occurrence, is of great importance. A distinction is made here between weak and strong co-occurrence, which in each case affects the human interoperability of the topics. We speak of weak co-occurrence when a word appears in many different documents with a high number of different words. In such a case, the probability that this word can represent a topic well decreases. As an example, the word "introduction" can be mentioned since this word is often used as a heading in many different documents. A strong co-occurrence exists when a word often occurs in combination with the same words. This increases the probability that the word represents a certain topic. For example, the words "Bourdieu" and "field" or "Bourdieu" and "practice" can be weighted higher in a common topic and thus, a meaningful interpretation of the topic is possible.

LDA algorithm includes several important parameters, such as distribution parameters which influence the probability distributions already described (Blei 2012). One of the most important parameters is the number of topics. If too few topics are determined, it is possible that important topics are not determined or that different topics are combined in one topic. If a too high number of topics is determined, this happens at the expense of the interpretability of the topics since otherwise too many different words are combined and get a high load.

A frequently used metric is the coherence measure C by Mimno et al. (2011). The higher this measure, the better the topics should be interpretable. Still, this does not mean that human interpretation is irrelevant. In fact, Mimno et al. (2011) argue in

favour of using both the coherence measure and human interpretation as quality criteria for topic evaluation. For human interpretation to be successful, the researcher must have domain-specific knowledge, if not expert knowledge, to evaluate the topics. Roberts et al. (2014) emphasise that the coherence measure alone provides an inadequate evaluation of the number of Topics. The measure emphasizes the internal coherence of the words in the topics, while the distinction of topics by certain words among themselves, even of very similar topics, is neglected. In addition, to enable a more sophisticated evaluation, they used the excusivity of Topics (Bischof and Airoldi 2012).A word is defined as exclusive if it has a high probability for one topic and a low probability for other topics. We also use both measures in this study. After several models had good results for both metrics, the word lists of these models were intellectually compared. It was found that the model with 50 topics had the comparatively best word lists for interpretation. Topics that could not be interpreted were excluded.

Structural Topic Modelling (STM) essentially differs from the traditional LDA model as covariates can be integrated into the model, which influences how the distributions of the words within the topics as well as the distribution of the topics in the documents are calculated (see Roberts et al. 2014: 1067). Also, the STM Package in R provides further methods to analyse the results of the topic modelling (Rodriguez and Storer 2020). For this study, we performed a polynomial regression, which can be used to determine the probability of a topic occurring in documents over time. This method is implemented in the R STM package. To be able to investigate the cyclical nature of a topic over time, a polynomial regression is calculated for each topic. Polynomial regression is suitable if the relationship between an independent variable and a dependent variable is non-linear. The model can then be improved with a quadratic or higher-order term of x. As higher the potency, the regression line is mapped closer and closer to the data points. In the STM Package, the polynomial regression is modelled with a second-degree polynomial. In the context of topic modelling, the dependent variable is the probability of occurrence of a topic in a document. The independent variable is the date of publication of the document on the homepage of the party websites.

For a meaningful interpretation of the topics, both domain knowledge and the reading of relevant texts are helpful. LDA makes it possible to specify relevant texts. Based on the probability that a topic is contained in a document, relevant texts of the corpus can thus be identified. With this in mind, we used the five percent of documents with the highest probability and limited ourselves to checking these texts intellectually. Our aim was to be able to interpret the generated word lists better. For this purpose, the procedure was sufficient, and we dispensed with decided methods of qualitative text analysis.

4 Results

A total of 681 documents from eleven party websites were extracted from the WARC files for the period from 23 April to 9 June 2019.[4] The corpus contains a total of 670,125 tokens.[5] Table 1 shows the number of documents per party web page, as well as the relative

[4] The party "Die Partei" was removed from the corpus because the crawl retrieved only very few documents.

[5] The term token in Natural Language Processing refers to individual instances of words (as supposed to types that describe general classes).

Table 1 Documents per party website. Own depiction

Party website	N (documents)	Relative frequency of documents in percent	N (tokens)	Relative frequency of tokens in percent
www.afd.de	32	4.70	28,218	4.19
www.cdu.de	40	5.87	40,314	5.99
www.csu.de	113	16.59	104,884	15.59
www.die-linke.de	160	23.49	163,565	24.31
www.fdp.de	93	13.66	93,926	13.96
www.freiewaehler.eu	15	2.20	14,303	2.13
www.gruene.de	67	9.84	72,323	10.75
www.npd.de	38	5.58	36,262	5.39
www.oedp.de	32	4.70	28,718	4.27
www.piratenpartei.de	64	9.40	61,437	9.13
www.spd.de	27	3.96	26,175	3.89

number of documents in relation to the total corpus and the number of tokens per party and their relative frequency.

We identify variance between the different party websites. Most tokens (relative frequency of tokens) come from the websites www.die-linke.de (24.31%), www.fdp.de (13.96%) and www.gruene.de (10.75%). The number of documents is also distributed rather unevenly, with a range from 15 to 160 documents. In this regard, especially www.die-linke.de (160), www.csu.de (113) and www.freiewaehler.eu (15) stand out. This inequality is likely to affect the identification of the most common topics, which must be considered when interpreting the results. It must also be noted at this point that the mere identification of topics does not reveal whether the communication on the party websites is of a general nature or specifically part of the online campaign for the European elections. With regard to campaigning, one explanation why the website of the Freie Wähler was attributed a significantly lower number of documents could be that it has a comparatively weak organisation and is more of a regional or local party without major political aspirations on European level. The high number of documents on the CSU and Linke websites could indicate a particularly high level of political communication and campaigning activity via the medium of the website. Nevertheless, all parties seem to use their websites regularly as a communication channel. Apart from the website of Freie Wähler, we count an average of one post per day. Therefore, except for the outlier, all parties use the medium regularly.

Figure 2 shows the ten most frequent topics and their probability of appearance in the corpus. Since this case study serves as an example of the use of web archives in analysing communication on party websites during the 2019 European parliament election campaign, we will limit the presentation of results to the five topics that were most strongly represented in the course. The terminology and the choice of labels for the topics will be discussed, as well as their usage over time. The five most frequent topics in the corpus were awarded to the labels “Söder & Taxes & Redistributive Policies”, “Weber & European Policies”, “Dresden & Migration”, “Trade Unions & Linke” and “Beer European Election Campaign”.

Taking a first look at the top five topics, we get an overview of a wide range of issues. It is striking that three of the five topics contain politicians’ names. Two of them—Weber and Beer—most probably refer to the German and European top candidate for the European

Table 2 Top five topics in order of their probability. Own depiction

Label	terms
Söder, Taxes & Redistributive Policies	söder, steuererhöhungen, grundrente, csu_chef, grundsteuer, bundesfinanzminister, csu-vorsitzende, markus_söder, bayerisch, heil [Söder (Bavarian PM), Tax increase, Basic Pension, Head of CSU, Tax on Land and Buildings, Federal Minister of Finance, Chairmen of the CSU, Markus_Söder, bavarian, Heil (Federal Minister of Employment, SPD)]
Weber & European Policies	weber, krebs, manfred_weber, söder, kommissionspräsident, europa, blume, sozialist, europäer, türkei [Weber, Krebs, Manfred Weber, Söder, President of the European Commission, Europe, Blume, Socialist, Europeans, Turkey]
Dresden & Migration	dresden, töten, migranten, migration, plakat, richter, franzen, stadt, ausstrahlen, peter [Dresden, kill, Migrants, Migration, Bill, (Peter) Richter, Franzen, City, broadcast, Peter (Richter)]
Trade Unions & Left	teilnahme, gewerkschaft, genosse, gewerkschaftlich, genossin, linke, linken, veranstaltung, beschäftigte, bundesgeschäftsstelle [Participation, Trade Union, Comrade (male), Union (adjective), Comrad (female), Linke (party), Left, Event, Employees, National Party Headquarter]
Beer European Election Campaign	beer, kulturell, europäische_union, nicola_beer, brexit, macron, abstimmung, europa, bürgerin, brüssel [Beer, Cultural, European Union, Nicola_Beer, Brexit, Macron, Vote, Europe, Citizen (female), Brussels]

elections, Manfred Weber, CSU/EPP-parliamentary group, and the German top candidate Nicola Beer, German liberal party (FDP). The third name refers to Markus Söder, the CSU-party leader. Furthermore, different types of policies seem to play a role in the communication of the party websites. These contain taxes and redistributive policies as well as European policies. At this stage of the analysis our exploratory approach provides neither a clear indication of the role that the European Elections 2019 play in the communication on the party websites, nor a clear emphasis on either national or European political issues. We see that top candidates, European policies and European election campaigns are addressed. The topics "Trade Union and Left" and "Söder & Taxes & Redistributive Policies" cannot be clearly allocated at this stage of the analysis. The topic "Dresden & Migration" could, however, deal with national or a mixture of national and European issues. In order to get a deeper insight into which issues were of importance in the 2019 European election campaign of German parties, we take a closer look at the topic terms.

Table 2 lists the top five topics (in order of their probability) with the associated label and the ten words with the highest load for the respective topic. A closer look at the terms per topic underline the assumption that the topics vary and cannot be specifically assigned to national or European issues. We identify both national and European topics, as well as personalisation on a state, federal, and European level.

The list of terms for the most frequent topic "Söder, Taxes & Redistributive Policies" contains references to the CSU party leader and Bavarian Prime Minister Markus Söder (Söder (Bavarian PM), Markus_Söder, bavarian) and his position within the party (Chairmen of the CSU) as well as tax and fiscal policy issues. The individual reading of the 11 texts (nine from www.cdu.de, two from www.fdp.de), that was carried out to verify the results of the distant reading via topic modelling, showed that the topic assignment could

be confirmed in 7 of 11 cases. Furthermore, the texts can be ascribed to further categories like tax policy—explicitly property tax—financial policy, Germany, the CSU, pension policy—explicitly basic pension—and the SPD. The category European election campaign only occurred once.

Figure 3 shows the probability of appearance of the topic depending on the party website. The larger the circle, the higher the probability that the topic will appear on a particular party website. The 95% confidence interval is shown in red. It is not surprising that the topic "Söder & Steuern & redistributive Politik" is mainly found on the CSU website, since the party chairman is a central topic there. Since this topic appeared most frequently in the whole corpus and it was mainly discussed on one party website, this could on the one hand indicate that the CSU party communication on the website during campaign during the European parliament elections 2019 was very much focused on Söder, which is all the more remarkable considering the fact that he was not even up for election. On the other hand, the CSU could also have published general issues of a domestic nature on its website rather than issues related to the European elections. As we also identified the topic "Weber & European Policies", this finding could lead to the assumption, that there are two distinct tendencies – one with regard to the party chairman and the other with regard to the Spitzenkandidat—in the CSU election campaign.

Figure 4 illustrates the development of the topic over time. This function shows the probability of the topic depending on the timestamp of the document on the party website. The top and bottom lines show the 95% confidence interval. The topic was more prominent in the first third of the investigation period. Towards the date of the European election, it remained at a relatively low level. This puts the previous findings into perspective: although the topic on Markus Söder and taxes occurs frequently, the probability of its occurrence decreases around the date of the election. If one assumes that communication on the party website in the immediate period around the election tends to revolve around this event, the party leader seems to play a less important role. At the end of the period, the 95% confidence interval becomes very large, which is probably due to a poorer data basis. It is therefore not possible to give a more detailed interpretation of this period. It can therefore be assumed that this nationally oriented topic occurs frequently in general communication, but sees a decrease during the elections for the European Parliament. As such, it is probably not as prominent in communications leading up to this specific event.

The second most frequent topic "Weber & European Policies" contains the terms Weber, Krebs, Manfred Weber, Söder, President of the European Commission, Europe, Blume, Socialist, Europeans, Turkey. The list of words refers to the European People's Party (EPP) group Spitzenkandidat Manfred Weber very often (Weber, Manfred_Weber, President of the European Commission). In addition, Markus Blume (CSU general secretary) and Markus Söder, and terms referring to Webers campaigning issues of Europe and its political relationship with Turkey, are mentioned in relation to this topic. An individual reading was performed on 15 texts (11 from cdu.de, one from csu.de, three from www.gruene.de). The topic assignment could be fully affirmed in 7 of 15 cases. In three cases, Weber played no role, and two times the topic assignment could not be affirmed at all. Through reading, issues corresponding with the categories European election campaign, CDU, CSU, EU, Robert Schumann, Germany and France were identified. This underlines that the second most frequent topic contains election campaign content of the CDU/CSU with reference to their top candidate Manfred Weber.

Figure 5 shows the probability of the topic occuring on a specific website. A clear emphasis on the topic can be identified for one party website as the topic appears mainly

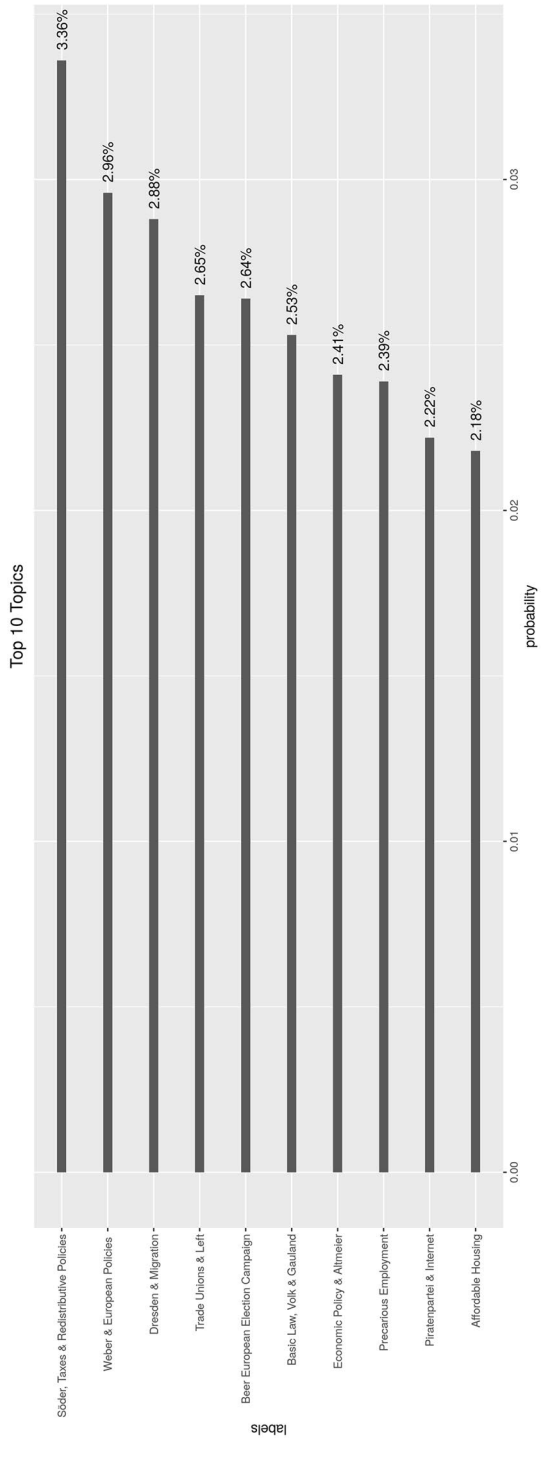

Fig. 2 Top 10 topics. Own depiction

Fig. 3 Topic Söder, Taxes & redistributive policies. Own depiction

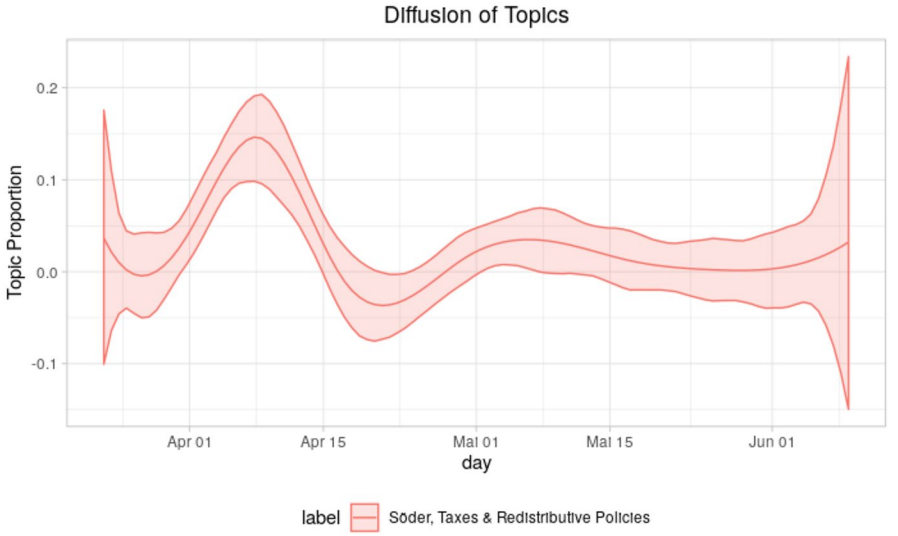

Fig. 4 Topic Söder, Taxes & Redistributive policies—development over time. Own depiction

on the website of the CDU. On the website of Weber's Party CSU, however, the Söder-topic prevails. This suggests that the CSU published Söder-related texts rather than content concerning the European elections and Spitzenkandidat and the parallelism described above is shifting in favour of Söder-related content.

The third most frequent topic "Dresden & Migration" (Dresden, killed, Migrants, Migration, Bill, (Peter) Richter, Franzen, City, broadcast, Peter (Richter)) appears to be a national topic at first glance. However, an individual reading of the 19 texts (17 from www.npd.de, two from www.afd.de, two from www.die-linke.de) revealed a close relation to the European parliament elections as the texts from the NPD addressed the migration issue. All texts from the AfD and the NPD can clearly be assigned to the

Fig. 5 Topic Weber & European policies. Own depiction

Migration topic, the two texts from the website www.die-linke.de do not fit the topic. The frequent occurrence of "Dresden" could be due to the fact that these are mainly texts from the NPD in the state of Saxony, whose capital is Dresden. In the reading, the further categories of “violence by refugees”, “European election campaign” (including “election posters” and “election commercials”), “administrative court”, “public broadcasting”, “NPD” and “racism” could be identified. The close reading shows that the topic could be related to election campaign-oriented communication by German populist parties because they tend to harbour strong feelings towards immigration.

The fourth topic “Trade Unions & Linke” contains mainly terms that are to be understood in the context of trade unions and their work, such as "Trade Union", "Comrade" (both male and female forms appear separately), "Participation", "Event". The german word "linke" could refer to a political camp, or to the party "Die Linke". Fig. 6, shows that "linke" probably refers to the German party of the same name, as the topic is clearly visible on their party website. As can be expected, the topic was particularly present before 1 May (Labour Day). Reading the 20 texts (18 from www.die-linke.de, 1 from www.oedp.de, 1 from www.spd.de), eight can be directly assigned to the topic. Six texts deal only with the party Die Linke, three only with trade unions and the texts on www.oedp.de and www.spd.de do not correspond. Further categories that can be ascribed to the topic are "Racism" or "Against Racism", "European Election Campaign", "Social Policy", "EU Problem Fields", "EU Reforms" and "Party Events”. It is not unusual that parties deal with other issues during election campaigns.

The last prominent and highly distinct topic is labelled the "Beer European Election Campaign" because it fundamentally reflects the campaigning of the leading candidate of the German liberal party (FDP), Nicola Beer. The list of words contains terms that are most likely related to her election campaign (Brexit, Macron, Citizen (female)). This assumption is corroborated by correlating the topic to party websites. As Fig. 7 illustrates, the topic correlates highly to the website of the FDP. Remarkably, yet another party website, namely that of the Freie Wähler, correlates closely with the set of topics attributed to the FDP. Reading the 22 texts (20 from www.fdp.de, 1 from www.gruene.de, 1 from www.oedp.de), 20 texts from the FDP website can be directly assigned to the topic. The other two texts can also be assigned to the European election campaign,

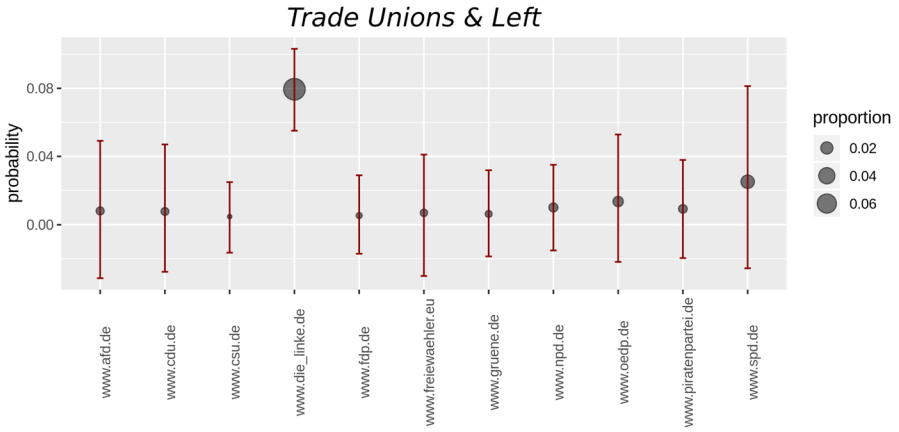

Fig. 6 Topic trade unions & left. Own depiction

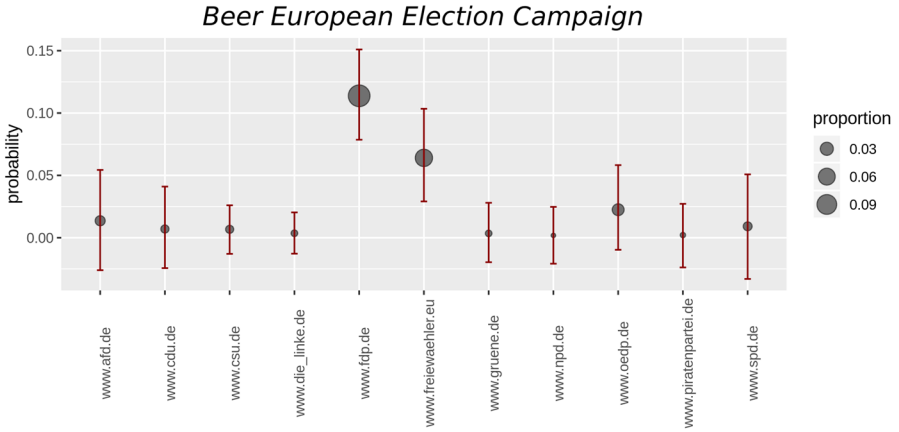

Fig. 7 Topic beer European election campaign. Own depiction

but to the Green Party or the ÖDP. This indicates that the communication and the campaigns relate to similar EU issues and that there are debates between the parties.

5 Discussion

The results of our analysis show that applying granular extraction of unstructured temporal data from web archives and structuring this content for analysis with topic modelling in mixed methods research designs is a useful approach to dealing with challenges arising from the shift of election campaigns and political discourses to digital or virtual arenas. Since collecting, storing and archiving this type of data cannot reliably be covered by traditional methods, by archiving websites, researchers are enabled to both document and analyse the shifts in communication on the internet.

Furthermore, research on the tracing of political and social debates cannot be conducted without a closer investigation of the party's and politician's websites and social media platforms. Although our analysis focused on party websites and therefore a top-down information strategy of online campaigns without considering social media platforms, this limited corpus can be considered as adequate for testing solutions to the methodological challenges. Against the backdrop of research on the second-order character of European elections, we used an exploratory approach to detect which issues were of importance in the 2019 European election campaign. We identified both national and European issues and a high level of personalisation, especially with regard to top candidates of the European elections 2019. However, it seems likely that policy issues are not captured by the method because they are very diverse in content and thus also in the language that describes them. While the names of top candidates are easy to identify, picking up policy issues is associated with different terminologies, which in turn coincide between different parties. In our analysis, we could not discern in how far campaigns were dominated by national or European political topics. Instead, both national and European issues occurred. With regard to the distinction between first- and second-order elections, this result clearly confirmed the results of other studies of the European elections 2019 (e.g. Ehin and Talving 2021; Ehin and Talving 2021; Anders and Träger 2019; Pasquino and Marco Valbruzzi 2019; Plescia et al. 2019; De Sio et al. 2019; Haßler et al. 2019a, b), which attribute at least a partial second-order character to the European election. Furthermore, our results show that the investigated parties used their websites in very different ways. Quite surprisingly, the two parties with the highest number of publications were the conservative Bavarian regional party CSU and "die Linke", two parties one usually would not expect to be the frontrunners in digital campaigning. On the contrary, the populist party AfD belonged to the parties with the lowest web activity, a result that does not fit the internet savvy image of populist parties (Schaub and Morisi 2019). However, embedded social media elements that frequently appeared on the website of the Bündnis 90/Die Grünen, for example, could not be considered when extracting data from the Webarchives with warc2corpus. Including this type of data could have led to different results. Although this analysis was conducted on the basis of a limited data corpus, it questions previous findings and could thus serve as a starting point for future analysis in which different types of web data like party websites, candidate websites and their specific social media accounts (Twitter, Instagram, Facebook) could be harvested and archived separately with standardised methods and analysed with topic modelling. Although the results are only preliminary and based on a restricted data set, they show that analysing web pages by using web archives and distant-reading methods can bring new insights into current academic debates.

6 Conclusion

In our analysis with methodological focus, we aimed to show the benefits of creating web archives as a standardised technique for preserving online content, processing this type of unstructured data and applying topic modelling to explore and analyse it with regard to a research question. We therefore focussed our case study on the web campaigns of German political parties in the run up for the European Parliament elections 2019. We applied topic modelling as distant reading methods on web archives, to get an insight in the communication on the party websites and investigate which topics emerged frequently during the 2019

European election campaign. The research question was limited in scope but sufficient to test our methodological toolkit.

Therefore, particular focus was put on the challenges of using web archives and the warc2corpus software, which was developed in the context of our research project and facilitates for a finely grained extraction of data from WARC files. We illustrated the explanatory power of combining web archiving with computational approaches to text analysis on a research question from the field of European integration research. To eliminate the problems of using existing web archives, which are usually based on an unknown system and methodology of data collection, we decided to build our own web-archive. In doing so, we created an event crawl-based web archive which was limited in scope but sufficient to test our methodological toolkit on the aforementioned research question.

Since the focus of the analysis was on the methodological challenges, the results of the case study only give a first impression of what can be achieved with this new methodology. Regarding the dataset, our approach revealed that specifically compiled corpora are particularly suitable for analysing specific questions. Thus, we built our web archive through event crawls with web harvesting tools, based on an actor-centred seed list. However, corpus creation is only the first step in the analysis as web archiving per se only serves the purpose of reliable data acquisition and storage. To comprehensively analyse the information contained in web archives, advanced methods of information retrieval and data analysis are required. Because the basis for distant reading methods are raw text data, we needed to find a way to extract relevant text data from the set of WARC files in our archive and enrich it with metadata like party and date. We developed an open-source tool named warc2corpus which extracts precise, structured information at the cost of additional configuration. Applying W2C is a suitable way of accessing the required information in the WARCs and building a text corpus, but its application is demanding. In total, our corpus consisted of 681 documents from eleven party websites covering the period from 23 April to 9 June 2019. Regarding the composition of the corpus, initial observations could be made before topic modelling was applied. We identified variance between the different party websites, which suggests that the websites are used differently for election campaigns. However, it has to be considered that at this stage of the methodological development embedded social media elements that frequently appeared on the website of Bündnis 90/Die Grünen, for example, could not be extracted with W2C. Considering the size of the corpus, it seems appropriate to use the method of topic modelling which is already widely used in disciplines like communication studies, literature science and political science. For the purposes of this paper, we used Structural Topic Modelling (STM), which enabled us to investigate topics in relation to variables external to the text as such, and derived from the web archive, such as date and party. From a subject-specific perspective, we applied topic modelling to investigate whether the web pages of the selected political parties do still reflect the image of European elections as second-order elections or whether we can find characteristics of first-order elections. The results show that distant reading techniques like STM allow us to explore and analyse the web presence of political parties without reading every single web page. This is of particular importance for further research with web archives, especially if the amount of data exceeds the size that can be handled by traditional close reading methods. Even though in our case, the data set was limited in size, the findings are transferrable to larger corpora since our techniques of making data usable for the application of digital humanity methods (i.e. particularly warc2corpus) as well as the distant reading methods itself do not require a limitation of the analysed data.

However, without a closer look at the text data, results must be treated with caution and findings should be verified with traditional methods. Referring to Lemke et al. (2015), we

support the quest for an "integrated concept [of distant and close reading] that balances the advantages and disadvantages of the two concepts" (Lemke et al. 2015: 8), labelled in the literature as blended reading. Blended reading joins distant reading techniques with the intellectual examination and contextualisation of the results. To apply this method to such a corpus, however, requires both improvements to our instruments and our methodological apparatus and gives rise to questions of representativeness and inclusiveness. This could therefore be a topic for future research.

Authors' contributions Not applicable.

Funding Open Access funding enabled and organized by Projekt DEAL. Deutsche Forschungsgemeinschaft (DFG)—project number: 395175156.

Availability of data and materials According to §60d UrhG (German Act on Copyright and Related Rights) the data can only be made available to the public for a specifically limited circle of persons for their joint scientific research, as well as to individual third persons for the purpose of monitoring the quality of scientific research. The Bavarian State Library will store the data with appropriate measures against unauthorised use by third parties as long as it is necessary for scientific purposes or validation of scientific data.

Code availability https://github.com/sepastian/warc2corpus.

Declarations

Conflict of interest Not applicable.

References

Alvarez, R.M.: Computational Social Science: Discovery and Predictions. Analytical Methods for Social Research. Cambridge University Press, New York (2016)

Anders, L.H., Träger, H.: Die Europawahl 2019 – wieder eine second-order election? Eine Analyse der Wahlergebnisse in den 28 EU-Staaten. In: Kaeding, M., Müller, M., Schmälter, J. (eds.) Die Europawahl 2019. Ringen um die Zukunft Europas, pp. 315–326. Springer, Wiesbaden (2020)

Archives Unleashed Project: Archives Unleashed Toolkit (Version 0.50.0). Apache License, Version 2.0 (2019)

Barberà, P., Casas, A., Nagler, N., Egan, P.J., Bonneau, R., Jost, J.T., Tucker, J.A.: Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. Am. Polit. Sci. Rev. **113**(4), 883–901 (2019)

Bentivegna, S., Marchetti, R.: Campaigning on Twitter: The use of social media in the 2014 European elections in Italy. In: Davis, R., Holtz Bacha, C., Just, M.R. (eds.) Twitter and Elections Around the World. Campaigning in 140 Characters or Less, pp. 126–140. Routledge, New York (2017)

Bischof, J. M., Airoldi, E. M.: Summarizing topical content with word frequency and exclusivity. In: Proceedings of the 29th International Conference on Machine Learning (ICML'12), pp. 9–16. Omnipress, Edinburgh (2012)

Blei, D.M.: Probabilistic topic modeling. Commun. ACM **55**(4), 77–84 (2012)

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Blind, J.: Das Heimspiel der „Europa-Parteien“? Die Europawahlkämpfe der Union von 1979 bis 2009. Springer, Wiesbaden (2012)
Braun, D., Tausendpfund, M.: Die neunten Direktwahlen zum Europäischen Parlament. Rahmenbedingungen, Parteien und Bürger in der Bundesrepublik Deutschland. Zeitschrift für Parlamentsfragen 50/4, pp. 715–735 (2019)
Braun, D., Schmitt, H.: Different emphases, same positions? The election manifestos of political parties in the EU multilevel electoral system compared. Party Polit. **26**(5), 640–650 (2020)
Braun, D., Schwarzbözl, T.: Put in the spotlight or largely ignored? Emphasis on the Spitzenkandidaten by political parties in their online campaigns for European elections. J. Eur. Public Policy **26**(3), 428–445 (2019)
Brügger, N.: Web archiving—between past, present, and future. In: Consalvo, M., Ess, C. (eds.) The Handbook of Internet Studies, pp. 24–42. Wiley, Chichester (2011)
Brügger, N., Milligan, I. (eds.): SAGE, Los Angeles (2018T)
Brügger, N., Milligan, I.: Introduction: internet histories and computational methods. Intern. Hist. **3**(3–4), 199–201 (2019)
Brynjolfsson, E., McAfee, A.: The second machine age: work, progress, and prosperity in a time of brilliant technologies. Norton, New York (2014)
Buriol, L.S., Castillo, C., Donato, D., Leonardi, S., Millozzi, S.: Temporal analysis of the Wikigraph. Web Intelligence, pp. 45–51 (2006)
De Sio, L.: Explaining the outcome. Second-order factors still matter, but with an exceptional turnout increase. In: De Sio, S., Franklin, M.N., Russo, L. (eds.) The European Parliament Elections of 2019, pp. 57–66. University Press, New York (2019)
De Vreese, C.H.: Second-rate election campaigning? An analysis of campaign styles in European parliamentary elections. J. Polit. Market. **8**(1), 7–19 (2009)
Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
Ehin, P., Talving, L.: Second-order effects or ideational rifts? Explaining outcomes of European elections in an era of populist politics. Ital. Polit. Sci. Rev. **50**(3), 350–367 (2020)
Ehin, P., Talving, L.: Still second-order? European elections in the era of populism, extremism, and Euroscepticism. Politics **41**(4), 467–485 (2021)
Jannidis, F., Kohle, H., Rehbein, M. (eds.): Digital Humanities: Eine Einführung. Metzler, Weimar (2017)
Fuchs, M., Holnburger, J.: #ep2019 – Die digitalen Parteistrategien zur Europawahl 2019. Friedrich-Ebert-Stiftung, Hamburg/Berlin (2019)
Galpin, C., Trenz, H.J.: Rethinking first- and second-order elections. media negativity and polity contestation during the 2014 European Parliament Elections in Germany and the UK. ARENA working paper 3/2018 (2018)
Galpin, C., Trenz, H.J.: In the shadow of Brexit: the 2019 European parliament elections as first-order polity elections? The Political Quarterly 90/4, pp. 664–671 (2019)
Gassner, S.: sepastian/warc2corpus. GitHub. https://github.com/sepastian/warc2corpus (2020). Accessed 23 May 2021
Gattermann, K., De Vreese, C.H., van der Brug, W.: No longer second-order? Explaining the European Parliament elections of 2019. Politics **41**(4), 423–432 (2021)
Göler, D., Reiter, F.: “Let´s archive!” Die Dokumentation internetbasierter Daten als neue Herausforderung für die europäische Integrationsforschung. Integration **42**(4), 321–328 (2019)
Göler, D.: Endlich ein echtes Parlament? Die Rahmenbedingungen des Vertrages von Lissabon und das Europäische Parlament. In: Mittag, J. (Ed.): 30 Jahre Direktwahl zum Europäischen Parlament (1979–2009). Europawahlen und EP in der Analyse, pp. 289–311. Nomos, Baden-Baden (2011)
Grill, C.: How Anti-european, populist parties campaigned in the 2014 EP election. In: Holtz-Bracha, C. (ed.) Europawahlkampf 2014, Internationale Studie zur Rolle der Medien, pp. 75–96. Springer, Wiesbaden (2016)
Grimmer, J., Stewart, B.: Text as Data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. **21**, 267–297 (2013)
Haßler, J., Magin, M., Russmann, U., Baranowski, P., Bene, M., Schlosser, K., Wurst, A-K., Fenoll, V., Kruschinski, S., Maurer, P.: Reaching Out to the Europeans. Political Parties‘ Facebook Strategies of Issue Ownership and the Second-Order Character of European Election Campaigns. In: Zur Rolle der Medien., Holtz-Bacha, C. (eds.) Europawahlkampf 2019a, pp. 87–114, Springer, Wiesbaden (2020)
Haßler, J., Magin, M, Russmann, U., Fenoll, V.: Campaigning on Facebook in the 2019b European Parliament Election. Informing, Interacting with, and Mobilising Voters, Plgrave Macmillan, London (2021)

Haußner, S., Klika, C: Who ist connected to whom? A twitter-based network analysis of members of the European parliament elected in 2019. In: Kaeding, M., Müller, M, Schmälter, J. (eds.) Die Europawahl 2019. Ringen um die Zukunft Europas, pp. 423–436, Springer, Wiesbaden (2020)
Haußner, S., Kaeding, M.: Je höher, desto besser? – Die Europawahl 2019 vor dem Hintergrund sozial verzerrter Wahlbeteiligung. In: Kaeding, M., Müller, M., Schmälter, J. (eds.) Die Europawahl 2019. Ringen um die Zukunft Europas, pp. 327–340. Springer, Wiesbaden (2020)
Hix, S., Marsh, M.: Second-order effects plus Pan-European political swings. An analysis of European parliament elections across time. Electoral Stud. **30**, 4–15 (2011)
Hobolt, S., Wittrock, J.: The second-order election model revisited: An experimental test of vote choices in European Parliament elections. Electr. Stud. **30**, 29–40 (2011)
Hockey, S.: Electronic Texts in the Humanities. Principles and Practice. Oxford University Press, Oxford (2000)
Höller, I.: Haupt- und Nebenwahlkämpfe?: Mediale Berichterstattung und politische PR in österreichischen Wahlkämpfen. LIT Verlag, Münster (2015)
Holtz-Bacha, C.: Europawahlkampf 2019. Zur Rolle der Medien. Springer, Wiesbaden (2020)
Holtz-Bacha, C., Leidenberger, J.: Europawahl 2009: Wahlkampf im Schatten der Bundestagswahl oder doch eine europäische Kampagne? In: Holtz-Bacha, C. (ed.) Die Massenmedien im Wahlkampf – Das Wahljahr 2009, pp. 22–41. Springer, Wiesbaden (2010)
Holtz-Bacha, C.: Europawahl 2014. In: Holtz-Bacha, C. (ed.): Europawahlkampf 2014: Internationale Studien zur Rolle der Medien, pp. 1–13. Springer, Wiesbaden (2016)
Hooghe, L., Marks, G.: A postfunctionalist theory of european integration: from permissive consensus to constraining dissensus. Br. J. Polit. Sci. **39**(1), 1–23 (2009)
Hooghe, L., Marks, G.: Grand theories of European integration in the twenty-first century. J. Eur. Publ. Policy **26**(8), 1113–1133 (2019)
Irwin, G.: Second-order or third-rate? Issues in the campaign for the elections for the European Parliament 1994. Electr. Stud. **14**, 183–199 (1995)
ISO: ISO 28500:2017 Information and documentation—WARC file format. International Organization for Standardization (2017)
Kaiser, C.: Soziale Medien als Mittel der Produktgestaltung (Co-Creation). In: König, C., Stahl, M., Wiegand, E. (eds.) Soziale Medien. Gegenstand und Instrument der Forschung, pp. 171–194. Springer, Wiesbaden (2014)
Karpf, D.A.: Social science research methods in internet time. Inf. Commun. Soc. **15**(5), 639–661 (2012)
Knecht, S., Debre, M. J.: Die „digitale IO“: Chancen und Risiken von Online-Daten für die Forschung zu Internationalen Organisationen. ZIB Zeitschrift für Internationale Beziehungen **25**(1), 175–188 (2018)
Lemke, M., Niekler, A., Schaal, G., Wiedemann, G.: Content analysis between quality and quantity: fulfilling blended-reading requirements for the social sciences with a scalable text mining infrastructure. Datenbank Spektrum **15**(1), 7–14 (2015)
Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third International Conference on Web Search and Web Data Mining (WSDM), New York (2010)
Lilleker, D.G., Jackson, N.A.: Political Campaigning, Elections and the Internet. Routledge, London (2011)
Maggini, N.: Understanding the electoral rise of the five star movement in Italy. Czech J. Polit. Sci. **21**(1), 37–59 (2014)
Marchal, N., Kollanyi, B., Neudert, L.-M., Howard, P.N.: Junk news during the EU parliamentary elections: lessons from a seven-language study of Twitter and Facebook. Data memo 2019/3. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/EU-Data-Memo.pdf. (2019)
McFarland, D.A., Ramage, D., Chuang, J., Heer, J., Manning, C.D.: Differentiating language usage through topic models. Poetics **41**(6), 607–625 (2013)
Mimno, D., Wallach, H. M., Talley, E., Leenders, M., McCallum, A.: Optimizing Semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11), pp. 262–272 (2011)
Moretti, F.: Distant reading. Verso. London, New York (2013)
Niedermayer, O.: Von der "nationalen Nebenwahl" zur "europäisierten Wahl"? Die Wahl zum Europäischen Parlament vom 26. Mai 2019. Zeitschrift für Parlamentsfragen 50(4), 691–714 (2019)
Pasquino, G., Valbruzzi, M.: The 2019 European Elections: a 'second-order' vote with 'first-order' effects. J. Mod. Ital. Stud. **24**(5), 736–756 (2019)
Pfaffenberger, F., Heinrich, P.: Die überschätzte Gefahr? Twitter-Bots im Europawahlkampf 2019. In: Zur Rolle der Medien, Holtz-Bacha, C. (eds.) Europawahlkampf 2019, pp. 115–148, Springer, Wiesbaden (2020)

Plescia, C., Wilhelm, J., Kritzinger S.: First-order breakthrough or still second-order? An assessment of the 2019 EP elections. In: Plescia, C., Raube, K., Wilhelm, J., Wouters, J. (eds.) Assessing the 2019 European Parliament Elections, pp. 76–95, Routledge, London (2020)
Reif, K.: European elections as member state second-order elections revisited. Eur. J. Polit. Res. **31**, 115–124 (1997)
Reif, K., Schmitt, H.: Nine second-order national elections. A conceptional framework for the analysis of European election results. Eur. J. Polit. Res. **8**, 3–44 (1980)
Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Kushner Gadarian, S., Albertson, B., Rand, D.G.: Structural topic models for open-ended survey responses. Am. J. Polit. Sci. **58**(4), 1064–1082 (2014)
Rodriguez, M.Y., Storer, H.: A computational social science perspective on qualitative data exploration: using topic models for the descriptive analysis of social media data. J. Technol. Hum. Serv. **38**(1), 54–86 (2020)
Rußmann, U.: Webkampagnen im Vergleich von Bundestags- und Europawahl-kämpfen. In: Tenscher, J., Rußmann, U. (eds.) Vergleichende Wahlkampfforschung, pp. 55–74. Springer, Wiesbaden (2016)
Schaub, M., Morisi, D.: Voter mobilization in the echo chamber: broadband internet and the rise of populism in Europe. Eur. J. Polit. Res. **59**(4), 752–773 (2019)
Schmitt, H., Sanz, A., Braun, D., Teperoglou, E.: It all happens at once: understanding electoral behaviour in second-order elections. Polit. Gov. **8**(1), 6–18 (2020)
Schmitt, H., Toygür, İ.: European parliament elections of May 2014: driven by national politics or EU policy making? Polit. Gov. **4**(1), 167–181 (2016)
Schulte-Cloos, J.: Do European Parliament elections foster challenger parties' success on the national level? Eur. Union Polit. **19**(3), 408–426 (2018)
Stevenson, M., Ben-David, A.: Computational methods for web history. In: Brügger, N., Milligan, I. (eds.) The SAGE Handbook of Web History. SAGE, Los Angeles (2018)
Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Landauer, T. K., McNamara, D. S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis, pp. 427–448. Psychology Press, New York (2007)
Strohmaier, M., Zens, M.: Analyse Sozialer Medien an der Schnittstelle zwischen Informatik und Sozialwissenschaften. In: König, C., Stahl, M., Wiegand, E. (eds.) Soziale Medien. Gegenstand und Instrument der Forschung, pp. 73–95. Springer, Wiesbaden (2014)
Stulpe, A., Lemke, M.: Blended Reading. In: Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse. In: Lemke, M., Wiedemann, G. (eds.) Text Mining in den Sozialwissenschaften, pp. 17–61. Springer, Wiesbaden (2016)
Träger, H.: Die Europawahl 2014 als second-order election − Ein Blick in alle 28 EU-Staaten. In.: Kaeding, M., Switek, N. (eds.) Die Europawahl 2014. Spitzenkandidaten, Protestparteien, Nichtwähler, pp. 33–44. Springer, Wiesbaden (2015)
Valentini, C.: Social media use by main EU political parties during EP elections 2019. In: Bolin, N., Falasca, K., Grusell, M., Nord, L. (eds.) Euroflections: Leading Academics on the European Elections 2019, pp. 80–81. Mittuniversitetet, Sundsvall (2019)
Wilkerson, J., Casas, A.: Large-scale computerized text analysis in political science: opportunities and challenges. Annu. Rev. Polit. Sci. **20**, 529–544 (2017)
Winters, J.: Coda: Web archives for humanities research – some reflections. In: Brügger, N., Schroeder, R. (eds.) The Web as History. Using Web Archives to Understand the Past and the Present, pp. 238–248. UCL Press, London (2017)
Woyke, W.: Stichwort: Wahlen – Ein Ratgeber für Wähler, Wahlhelfer und Kandidaten. Springer, Wiesbaden (2013)
Zürn, M.: Politicization compared: at national, European, and global levels. J. Eur. Publ. Policy **26**(7), 977–995 (2019)