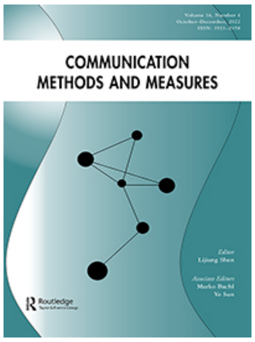

Communication Methods and Measures

ISSN: 1931-2458 (Print) 1931-2466 (Online) Journal homepage: www.tandfonline.com/journals/hcms20

Lifting the Veil on the Use of Big Data News Repositories: A Documentation and Critical Discussion of A Protest Event Analysis

Matthias Hoffmann, Felipe G. Santos, Christina Neumayer & Dan Mercea

COMMUNICATION METHODS AND MEASURES
2022, VOL. 16, NO. 4, 283–302
https://doi.org/10.1080/19312458.2022.2128099

Routledge
Taylor & Francis Group

Lifting the Veil on the Use of Big Data News Repositories: A Documentation and Critical Discussion of A Protest Event Analysis

Matthias Hoffmann[a], Felipe G. Santos[b], Christina Neumayer[a], and Dan Mercea[b]

[a]Department of Communication, University of Copenhagen - Københavns Universitet, Denmark; [b]Department of Sociology and Criminology, City, University of London, London, UK

ABSTRACT

This paper presents a critical discussion of the processing, reliability and implications of free big data repositories. We argue that big data is not only the starting point of scientific analyses but also the outcome of a long string of invisible or semi-visible tasks, often masked by the fetish of size that supposedly lends validity to big data. We unpack these notions by illustrating the process of extracting protest event data from the Global Database of Events, Language and Tone (GDELT) in six European countries over a period of seven years. To stand up to rigorous scientific scrutiny, we collected additional data by computational means and undertook large-scale neural-network translation tasks, dictionary-based content analyses, machine-learning classification tasks, and human coding. In a documentation and critical discussion of this process, we render visible opaque procedures that inevitably shape any dataset and show how this type of freely available datasets require significant additional resources of knowledge, labor, money, and computational power. We conclude that while these processes can ultimately yield more valid datasets, the supposedly free and ready-to-use big news data repositories should not be taken at face value.

Introduction

"Political protests have become more widespread and more frequent," is the headline of a newspaper article in *The Economist* published in 2020. The story refers to a report by the Washington-based think-tank Center of Strategic and International Studies (CSIS) (Haig, Schmidt, & Brannen, 2020, March 2). The research showcases an analysis run on data retrieved from the Global Database of Events, Language and Tone (GDELT). GDELT describes itself as a free and open platform, that:

> monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organisations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day (The GDELT Project, n.d.).

GDELT (Leetaru & Schrodt, 2013) indexes tens of thousands of multi-lingual news sites around the world and extracts information nearly instantaneously, from text, audio, and video formats, about events taking place around the globe. The project is the outcome of a collaboration between Google Jigsaw and George Mason University's Center for Social Complexity (Hopp et al., 2019). It promises nothing less than being "the largest, most comprehensive, and highest resolution open database of human society ever created" (The GDELT Project, n.d.). The global orientation of GDELT is eminent in their attempt to alleviate English language bias and a focus on the Western anglophone

CONTACT Felipe G. Santos f.santos@city.ac.uk University of London

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19312458.2022.2128099

part of the world. As the GDELT project states on its website: "we are incredibly excited by the ability of this high-resolution inventory, coupled with GDELT Translingual's ability to translate 98.4% of this material in realtime, to give voice to the most remote corners of the world in near-realtime" (The GDELT Project, 2015). With this promise, GDELT has widely been used as a data source in various media reports, policy documents, and academic research (e.g. Ferreira et al., 2021; Wang et al., 2016).

At first blush, such rich, feature-laden data turns GDELT into an exemplar "database of society," giving us access to insights we would otherwise be unable to uncover with tedious and time-consuming data collection and analysis processes. GDELT's potential for new insights is making an impact on time-consuming research areas such as communication research (Guo & Vargo, 2020; Hopp et al., 2020; Kwak & An, 2014; Yesilbas et al., 2021), studies of civil unrest (Levin et al., 2018; Ponticelli & Voth, 2020), and protest event analysis (Hutter, 2014), which may profit directly from automated analyses of large news data repositories. D'Ignazio and Klein (2020, p. 151), however, refer to such grand pledges as "Big Dick Data," foregrounding the masculinity of such fantasies inflating technical and scientific capabilities of automation and fetishizing size in favor of context. Communication scholars working with GDELT (Hopp et al., 2019) have pointed to limitations including little documentation about the process of collection and storage of data, complex and noisy data, and limited availability of tools that would allow the data to be used for scientific research. Other researchers have raised similar issues with automatically coded datasets such as the International Crisis Early Warning System (ICEWS, Jäger, 2018).

In this article, we chart the process of moving from the global news media repository to data that is amenable to a protest event analysis in six European countries. Within the wider scope of a project (ProDem, see: https://prodem.uni-frankfurt.de/) examining links between citizens, protests and parties and their follow-through into democratic policymaking, we envisaged using GDELT to compile a list of protests staged from 2015 to 2020 in six countries of interest indexed in the database, Denmark, Germany, Hungary, Italy, Romania, and the United Kingdom. We chose GDELT because the database includes a dedicated category for protests, and it covered the entire project timeframe. Additionally, GDELT promised a wealth of extra data on the events that it records. Our aim was to locate the largest mobilizations in each country, formulate a set of survey questions enquiring into respondents' participation in key protests that took place in those countries, and scrutinize cross-country and temporal variations in participation, as well as its links to other types of political engagement such as voting.

Hence, in this paper, we ask: What automation processes, material artifacts, and human labor are required for using GDELT in protest event analysis? By answering this question, the contribution of this article is twofold. From a methodological perspective, we suggest a way of maximizing the reliability of GDELT data for protest event analysis. Documenting the process ensures reliability (i.e. consistency and thus repricability of this research), therefore contributing to the results' validity (Krippendorff, 2018). From a conceptual perspective, we discuss the consequences that uncritical use of seemingly free big data repositories may have on the study of social phenomena. We contend that the issues we encountered are common to most research based on big data news repositories. With our work using GDELT, we submit that extracting information from big data news repositories requires a nuanced understanding of the materiality of the database or platform researchers collect data from as well as a non-negligible investment of labor in automated processes to analyze such data (see also, D'Ignazio & Klein, 2020). Drawing on a conceptual understanding that acknowledges the processual character of data and the invisibilities in such processes (e.g., Gitelman, 2013; Mattoni & Pavan, 2018; Neumayer, 2022), we argue that using the GDELT database for identifying protest events requires technical and scientific capabilities and labor that usually remain invisible. Moreover, we contend that the issues we encountered are common to most big data research.

In what follows, we review the literature, introduce the concepts connecting protest event analysis with critical data studies, and contextualize the use of GDELT for that purpose. We then discuss the process of employing GDELT for protest event analysis and the implications that surfaced through this research.

Studying the news: from hand coded data to machine learning

The study of news is one of the most fruitful lines of research in the social sciences. Communication scholars have a long tradition of understanding media's capacity to shape how the population perceives an issue (Entman, 1993; Lakoff, 1990; Scheufele, 1999), and which matters are considered relevant in social and political discourse (Dearing et al., 1996; McCombs & Shaw, 1972). Furthermore, social scientists drew on news reports to study protests (Jenkins & Perrow, 1977; McAdam, (1982/1999); Olzak, 1992), elections (Banks, 1997), and the evolution of public opinion (Galambos, 1975). Most of the research using news data in the past relied on labor-intensive human processing of the information contained in each piece. Recent computational advances have brought a considerable reduction of the money and time that are needed for this type of studies through the automation of these processes. Hence, there is a growing number of studies that rely on large scale, automatically coded datasets based on news data in fields such as communication studies (Hopp et al., 2020; Malik et al., 2021; Welbers et al., 2022), conflict research (Metternich et al., 2013), and protest event analysis (Bekker, 2022; Kurer et al., 2019).

While studies based on news data proved to be greatly productive, they have limitations, which are present both in manually coded as well as automatically processed datasets. For instance, even the most careful studies based on newspaper data will have to contend with the biases of queried news outlets (Earl et al., 2004). Two inter-related biases stem from the decisions by news outlets on what information to cover (selection bias), and from how outlets choose to characterize those events (description bias, Earl et al., 2004, p. 65). Techniques for mitigating those biases have been developed along with analytical procedures for the inclusion of sources guided by underpinning research questions, space and time boundaries, as well as other relevant contextual aspects (see also, Hutter, 2014).

Databases reliant on Machine Learning in general, and Natural Language Processing in particular are prone to biases (Hovy & Prabhumoye, 2021; Mehrabi et al., 2021). For our use-case of GDELT, this applies especially to the process of labeling events, entities, and locations. For example, if algorithms used to label locations or known entities are trained on annotated data that is biased toward major countries, we cannot expect them to perform as well on data from other parts of the world. Similarly, a changing repertoire of contention, i.e., "new" or formerly uncommon practices of protest might not be labeled a protest event if the classifier was trained on news reports on mass demonstrations from the 1960s and 70s. At the same time, we must consider the trade-off of more inclusive event definitions, that might introduce false positive results, and more restrictive ones, that might lead to an omission of false negatives. In either case, the decision to rely on databases such as GDELT means that researchers lack the resources to reproduce the impressive task of real-time news monitoring, translation, and event classification. Researchers ultimately subscribe to biases introduced by data repositories through their selection of data or their processing of data, but this must not keep them from critically assessing the quality of these data.

Invisibilities and the epistemology of big data news repositories

In the literature accompanying most big data repositories, these data are presented as a source of scientific truth imbued with objectivity that cannot be assumed for manual data collection. Unprecedentedly large quantities of data and computational power convey authority to research (D'Ignazio & Klein, 2020). Such discourses highlight objectivity and a "technological fix," signaling a "turn to AI" which extols big data's economic prospect and political power, now processed through artificial intelligence. All the while, they likewise promote the technological advancements and economic interests of commercial platforms that impel such developments through their funding schemes (Katzenbach, 2021).

Such rhetoric has been unmasked, for instance, by Wang et al. (2016), who point to problems with the validity and reliability of event data from free news repositories due to the complexity of coding interactions in news media – a task which involves processes such as actor recognition and normalization, geocoding, event encoding, timeframe detection, classification, or multilingual support. These authors acknowledge the potential of such data for giving insights into global problems, but this would require high-quality data, which is still not available. Thus, data quality is one crucial aspect that we need to consider when working with big data repositories. Moreover, we also need an epistemological "shift from administrative, positivist big data analytics" toward a critical approach "that combines critical social media theory, critical digital methods and critical-realist social media research ethics" (Fuchs, 2017, p. 47). Ultimately, even though Wang et al. (2016) focused their research on the Crisis Early Warning System (ICEWS), and the Global Data on Events Language and Tone (GDELT) datasets, which we consider in-depth here, similar issues have been raised about the Armed Conflict Location and Event Data Project (ACLED, Raleigh et al., 2010), the Uppsala Conflict Data Program Georeferenced Event Dataset (UCDP GED, Eck, 2012), and the Global Terrorism Database (GTD, LaFree, 2010).

The challenge at hand regards the reliability of insights gleaned from social media data or news media data repositories and their automated analysis (Diesner, 2015, p. 4). Data retrieved from big data news repositories is not an endpoint but rather a step in the process whereby researchers validate and transform data to turn it into reliable research data. This process (following Neumayer, 2022) starts with the creation of the data. In this case, the data are co-created by news media (deciding which protest events to cover) and the creators of the repository and, in some cases, their automated algorithms (deciding on the inclusion of media sources in their database). In the "datafication" phase, these news articles are curated, stored, labeled, and classified by the repositories' (often opaque) automated processing of data. In the final phase, researchers, journalists, and policymakers retrieve and analyze these data, again processing them with their own classification tools and methods and with a specific purpose. And made visible while others are rendered invisible, and some are quasi-invisible (Neumayer et al., 2021), as they are, for example, only known to the data repository. While we can make some of these processes visible (as we do in this research by tracing our process), these are mainly the decisions taken by us as researchers. Rendering opaque data-processing by big data repositories visible can be time- and labor-intensive or even impossible.

The use of GDELT in scientific research and beyond

GDELT monitors print, broadcast, and web news from across the world to extract events reported in them through Textual Analysis by Augmented Replacement Instructions (TABARI) and then code those events using Conflict and Mediation Event Observations (CAMEO), a well-established, hierarchical coding system for annotating event data typically retrieved from news sources (Best et al., 2013). Among other data, it identifies the type of event, its location, the initiator of the action, and its target (Leetaru & Schrodt, 2013). For socio-political events, GDELT indexes information ranging from statements by political leaders from across the world to various types of contentious collective action, violent and nonviolent events, any public activity by major international organizations and corporations, as well as other social groups. Moreover, GDELT uses additional software to extract the location of the event and the tone used in its news reporting (Leetaru & Schrodt, 2013, p. 18).

Given GDELT's global coverage of news reports spanning from 1979 to the present and the wealth of information it contains, it is no surprise that academics have used the dataset to study a great variety of topics. Studies using GDELT include communication-related research such as the spread of misinformation and fake news about COVID-19 (Bruns, Harrington et al., 2021; Bruns, Hurcombe et al., 2021); global news coverage of disasters and refugees (Kwak & An, 2014; Yesilbas et al., 2021); the influence of fake news on the online media ecosystem during the 2016 US Presidential elections (Guo & Vargo, 2020); the relation between the framing of news and socio-political events (Hopp et al., 2020); protest, revolutions, and other types of civil unrest (Christensen, 2019; Fengcai et al., 2020;

Levin et al., 2018; Ponticelli & Voth, 2020; Wu & Gerber, 2018) as well asand their repression by states (Christensen & Garfias, 2018); and responses to the COVID-19 pandemic by institutions and civil society (David Williams et al., 2021; Fu & Zhu, 2020; Yuen et al., 2021). Research drawing on GDELT data is published in some of the most reputed journals in the world, such as *Science* (Wang et al., 2016) and *Nature Scientific Reports* (Ferreira et al., 2021), and in the social sciences, such as *The Quarterly Journal of Economics* (Campante & Yanagizawa-Drott, 2018), *International Organization* (Christensen, 2019), and *Organization Sciences* (Odziemkowska & Henisz, 2021).

Beyond academic circles, media outlets, think tanks and policymakers base their analyses of protests and resistance on GDELT. As illustrated by our introductory quote, *The Economist* magazine claimed that "Political protests have become more widespread and more frequent" (The Economist, 2020). Similarly, the think tank The Carnegie Endowment for International Peace used it to report on the Arab Spring Revolution in Egypt (Austin Holmes & Baoumi, 2016); the popular political analysis site FiveThirtyEight, which is often featured in the New York Times[1] and ABC News,[2] used GDELT in a controversial piece claiming that "Kidnapping of girls in Nigeria is part of a worsening problem" (Chalabi, 2014); GDELT's founder, Kalev Leetaru, wrote a piece in *Foreign Policy* about the wave of protests that sparked the Arab Spring (Leetaru, 2014); the European Commission used GDELT data to model conflict events in "a conflict risk model supporting the design of the European Union's conflict prevention strategies" (Halkia et al., 2020); the US Army Engineer Research and Development Center (ERDC) lists GDELT among the datasets that can be used for "mission-relevant results" (Dos Santos et al., 2017, p. ii); and the United Kingdom's Office for National Statistics used GDELT to identify UK-based disasters and provides a comprehensive explanation of how to use GDELT (Williams, 2020).

GDELT and protest events

Protest event analysis (PEA) is a suite of methods used to collect and classify data about protests (e.g., their frequency and scope) from secondary sources, chief among which have been newspapers (Hutter, 2014). While most protest event analyses have relied on manual coding of newspaper data (e.g., Carvalho, 2022; Earl et al., 2003; Portos, 2021; Wang & Soule, 2012), scholars have started to use automated processes for producing these datasets.

A recent example of such a rigorously documented protest event database is the *Political Conflict in Europe in the Shadow of the Great Recession Protest Event Analysis* (POLCON_PEA), created by Kriesi et al. (2020). It comprises 30 European countries that were observed over a period of fifteen years (2000–2015). Through a combination of automation and manual annotation, its development maximized the availability of online news reports. Reflecting on this hybrid approach, the researchers stressed that while human manual coding represents the "gold standard" in PEA, the process becomes unfeasibly resource-intensive – as an expenditure of labor and time – as the number of counting units is scaled up into the thousands or millions (Kriesi et al., 2020, pp. 3–5). To minimize this cost, those researchers relied on a two-step method for unit selection which involved the initial use of Natural Language Processing techniques (including a domain-specific keyword search list) to retrieve news items for subsequent human classification. Yet even this scalable technique had to grapple with constraints, key among which was language heterogeneity. Faced with a universe of sources in multiple languages, the researchers chose to rely on the reporting in English of ten international news agencies indexed in the Lexis-Nexis database and which covered all the 30 countries. This protocol yielded north of five million news reports and was therefore described as "greedy" because of the number of false positive protest event matches that had to be sifted out thereafter.

[1] https://www.nytimes.com/2010/06/04/business/media/04silver.html.
[2] https://abcnews.go.com/US/fivethirtyeight-joins-abc-news/story?id=5453.

Protest event data from GDELT has received ample attention in academic research and beyond, yet its comparison with similar databases leads to contradictory results. Claassen and Gibson (2016) find a Spearman's rank-order correlation of 0.80 between GDELT and the Dynamics of Collective Action data (a hand-coded database using the New York Times as its source that is often considered as the gold standard in protest event datasets, see, Earl et al., 2003). Similarly, a comparison of daily counts of protest events during a period of high mobilization between November 2011 and 2012 in Egypt – in GDELT and ICEWS – found a correlation of r = 0.84 (Ward et al., 2013). However, studies comparing GDELT to three other computer-generated datasets that also monitor global societal events (ICEWS, GSR, and SPEED) found relatively poor correlations, below r = 0.3, in Latin America (Wang et al., 2016). Comparisons with other datasets led to similarly modest results (Hammond & Weidmann, 2014). This includes the Armed Conflict Location and Event Data Project mostly focused on the Global South (ACLED, Raleigh et al., 2010), and the Uppsala Conflict Data Program Georeferenced Event Dataset, which contains data about fatal violence taking place on the African continent (UCDP GED, Sundberg & Melander, 2013). Some of these discrepancies may originate in the different levels of press freedom in the regions compared. In countries where levels of freedom of press are lower, coverage of certain events is often limited (Drakos & Gofas, 2006), leading to a systematic under-reporting of certain types of activities by major outlets. In such cases, the fact that GDELT includes articles from news outlets reporting on events outside their country of origin may indeed result in a more accurate reporting of protest dynamics in certain countries.

Beyond the contradicting figures on how GDELT correlates with other datasets, there is a consensus about the number of false positives reported by the database. In the study by Wang et al. (2016) exploring the information contained in the URLs from which GDELT extracted data about protest events, only 21% of valid URLs (excluding duplicates) covered an actual protest event. There is likewise a lack of transparency on how GDELT includes sources over time, and it is difficult to distinguish the rereporting of historical events in the media from actual events. In comparison to ICEWS, GDELT consistently reports higher counts of protests in several countries and across different types of action, indicative of a higher number of false positives (*Ibid.*). While the problem of false positives is well known, researchers tend to use GDELT's data unfiltered, including in documents authored by European Institutions (Halkia et al., 2020), media outlets (The Economist, 2020) and financial corporations (Bolivar et al., 2021; Kolanovic & Krishnamachari, 2017; Ortiz & Rodrigo, 2018). Moreover, even though articles published in academic journals usually acknowledge GDELT's problem with false positives (Christensen & Garfias, 2018; Halkia et al., 2020; Yesilbas et al., 2021), they, nevertheless, often use the data unfiltered.

Various approaches have been used to mitigate the problem by filtering out false positives from the database. These include: eliminating events that do not appear in the first paragraph of the newswire, which GDELT identifies in the dataset with the binary variable 'root events'[3] (as events that appear in the first paragraph of news articles are likely to be central to the reporting in the article) combined with establishing a minimum number of sources from which the event was extracted (Claassen & Gibson, 2016; Consoli et al., 2021; Fu & Zhu, 2020; Odziemkowska & Henisz, 2021); using GDELT as a source of news items from a wide variety of outlets, accessing articles directly through the URLs provided by GDELT, and manually filtering and coding the information of interest for verification (Bruns, Harrington et al., 2021; David Williams et al., 2021; Vargo & Guo, 2017); unsupervised machine learning techniques to automatically classify GDELT events (Wright et al., 2020; Zheng, 2020); and assuming that events taking place in the same location during the same day are identical, thus reducing the number of detected events (Manacorda & Tesei, 2020). Common to all of these approaches is the fact that they are labor intensive and often designed to implement various mitigations, a behind-the-scenes process that is often rendered invisible when presenting the data.

[3] GDELT's understanding of what a "root event" is (events that appear in the first paragraph of an article) should not be confused with the meaning that some protest event analyst give to the concept (events that trigger other actions in a cycle of protests).

While most analyses relying on GDELT use the data rather uncritically and at face value, some scholars have evaluated GDELT's data and taken steps toward improving its quality. For instance, the iCore project[4] takes a critical stance toward the sources used by GDELT and restricts its queries to "a specifically constructed and extendable whitelist of 111 international, English-language major news outlets" (Hopp et al., 2019, p. 25),[5] providing key contextual information about each of them such as its country of origin or whether the outlet is government-owned (*Ibid.*). Beyond carefully selecting the highest quality news sources included in GDELT, iCore aims to place theory-driven research before isolated big data analyses by allowing researchers to filter the data by a number of topics.[6] Based on this approach, several studies have already used GDELT to analyze the connection between news frames and socio-political events (Hopp et al., 2020) as well as the use of moral language in communication about COVID-19 (Malik et al., 2021). In their turn, using a different approach, Manacorda and Tesei (2020) replicated the analysis they performed with GDELT with two other manually coded datasets with smaller reach as a robustness check.

Despite the great advances that iCore represents in making GDELT accessible to researchers, it still has limitations – some of them of its own and others derived from those of GDELT. First, at the time of writing this article, iCore only contains data from 2020 onwards,[7] a small portion of the time span covered by GDELT (from 1979,v.1; or 2015 onwards, v. 2). Second, while iCore provides a wealth of information about each data point, it excludes crucial data provided by GDELT for performing event analysis such as information about the initiator(s) and target(s) of the action, as well as the type of action performed and its location. In addition, iCore allows users to select data by topic, not by event type, making it unsuitable for the specific case of PEA. Third, the advanced methods used by the iCore team to process the data still require a sophisticated knowledge of computer technologies in order to comprehend how their final data was processed. Lastly, iCore provides only one url associated with the event (we assume it is the one identified by GDELT as "source url") but, as part of the noise that GDELT contains, sometimes this url is misleading as it has no information about the event actually coded by GDELT. While this option allows researchers to access data derived only from high quality sources without engaging in manual coding, researchers interested in event analysis would still need to access GDELT directly, potentially with a similar process to the one we describe below.

From event data to protest events: working with GDELT

The epistemological assumption that the GDELT database reports every protest event happening across the world rests on the notion that we understand the world through data. That presumption directs our attention to those events that are relevant enough to be studied (Mattoni & Pavan, 2018) by virtue of being represented by the GDELT data. Yet, in social research, we define data as aggregations of technologies (including data infrastructures, software, applications, code, machine learning, and AI), imaginaries prevalent in public discourse (e.g., the power conferred to big data might and their knowledge claims), and people (including data scientists, computer and social scientists, policymakers, media, business professionals or platform owners, Mattoni & Pavan, 2018, p. 314). Considering data within these three dimensions acknowledges the mystification of data and machine learning processes (e.g., Boyd & Crawford, 2012) and the way big data change our understanding of the world (Mayer-Schönberger & Cukier, 2013).

[4] https://icore.mnl.ucsb.edu/.

[5] While this step improves the quality of source-selection, it might again introduce a bias toward Western or English-language sources – a point GDELT originally sought to alleviate through a wide selection of sources. Our date retrieval through iCore however revealed that at the time of writing, the returned results included more sources than the whitelisted ones, which likely means that iCore is working toward including more balanced sources.

[6] For a full list of the topics covered by iCore, see: https://icore.mnl.ucsb.edu/event/.

[7] Through the peer-review process, we were informed that, at the time of writing this article, iCore was "undergoing a major architecture refactoring and "will soon again provide data from *February 2015* onward (as its initial, published release did), including data on events and event types (i.e. EventBaseCodes)". Our personal communication with the iCore team confirmed such plans but they do not address the full range of issues that we describe here.

With ICEWS and GDELT being the largest news data repositories available, the unprecedented "large amount of data" (Ferreira et al., 2021) is often used as a sole argument for the relevance and validity of scientific results. Taking into consideration that these data are not stable but always in process, we challenge such assumptions by considering the various steps involved in the process of making GDELT fit for research. In what follows, we reflect on the materiality of the data as well as the machine learning methods used to analyze them to suggest a way forward for research based on GDELT data. We unpack the process of using GDELT for protest event analysis, including a critical reflection on the challenges we encountered as well as suggesting one way of overcoming them.

Working with data from GDELT to identify protest events in the six European countries over a period from 2011 to 2021 was a long process with labor contributed by all coauthors at various stages. At an abstract level, the process represented a long string of decisions accounting for the characteristics of the data as predefined by the platform, often in opaque ways. While data we retrieve from such platforms is usually presented as unitary no matter how much manual and computational labor went into its development prior to its release, in this article we unpick this notion to manifest invisibilities (Neumayer et al., 2021) introduced when working with these data. Doing so allows us to understand better the materialities of data from large news media repositories, not as a constant but as a relational process (Mortensen, M. Neumayer & Poell, 2019). Those materialities are an amalgamation of the purpose of the project within which the data is used, the platform that sorts and curates the data, and the researchers who then employ these data for their analyses.

We divide this process into four phases that are indicative of the invisibilities we encounter when working with such data: a) extracting a protest database, b) addressing the false-positive problem, c) collecting more data and working with classifiers, and d) large-scale application of mitigations developed in previous stages. While interrelated, these phases do not always unfold in linear fashion. They are the product of incremental decisions made when working with the data (see, Figure 1, for an overview of the process). Pointing to these decisions challenges the epistemological assumption that large datasets can explain social phenomena in their own right and suggests that we need context-specific human expertise to interpret and process that data to render them valuable.

Extracting a protest database from GDELT: from events to mentions

Before data becomes available on GDELT, several processes have already been completed, i.e., collecting, curating, and classifying the data. Largely invisible, these processes have multiple implications, as we understood when attempting to extract a protest database suitable for our purpose. To compile a set of protest events, we queried GDELT[8] for all entries with *EventBaseCodes* 140, 141, 142, 143, 144, and 145. In the CAMEO taxonomy used by GDELT (Schrodt, 2012), these are defined as the following forms of action: "Engage in political dissent," "Demonstrate or rally," "Conduct hunger strike," "Conduct strike or boycott," "Obstruct passage, block," and "protest violently, riot." These categories can be broken down into four-digit codes in CAMEO depending on the specific form of action, but all are grouped under the "Protest" label because they constitute acts of contention. As we were interested in events that took place in Denmark, Germany, Hungary, Italy, Romania, and the UK, regardless of the nationalities of actors and targets, we used GDELT's *ActionGeo_CountryCode* variable to limit our query. We set a timeframe from January 2015 to December 2020. Although our period of investigation is 2011–2021, we had to limit ourselves to this time frame, as these are the earliest data available in GDELT Event Data Version 2. The differences in data structure between Versions 1 and 2 render a longitudinal study with the original timeframe challenging or even impossible.

At this early point in the process, we were aware that by using GDELT for PEA, we do not subscribe to an objective truth of "the protest event" but to a definition of protest imposed by CAMEO during

[8] A documentation of code and critical steps of our process is released on GitHub (see https://github.com/walfaelschung/GDELT_flow).

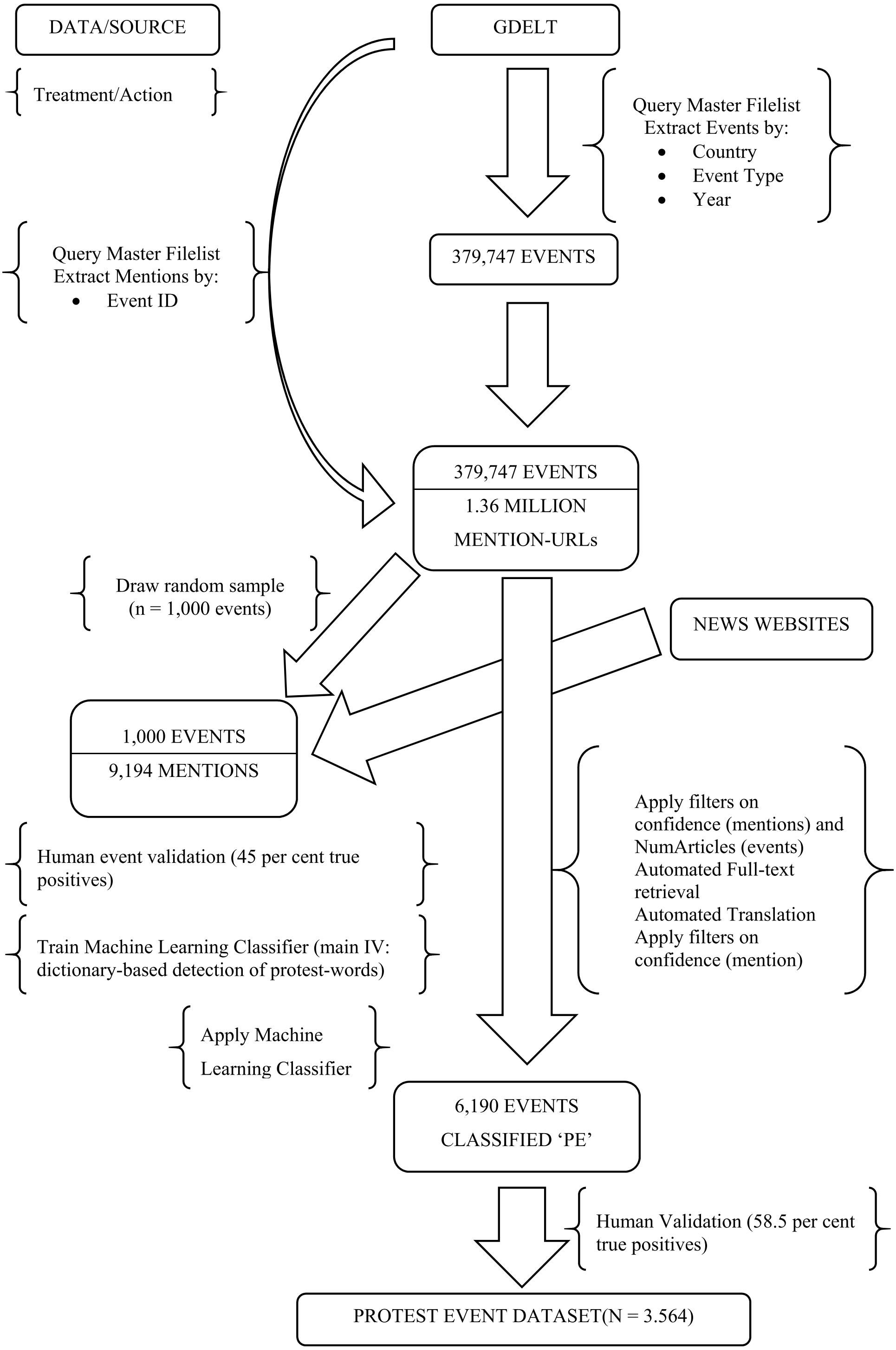
DATA/SOURCE
Treatment/Action
GDELT
Query Master Filelist Extract Events by:
• Country
• Event Type
• Year
379,747 EVENTS
Query Master Filelist Extract Mentions by:
• Event ID
379,747 EVENTS
1.36 MILLION MENTION-URLs
Draw random sample (n = 1,000 events)
NEWS WEBSITES
1,000 EVENTS
9,194 MENTIONS
Apply filters on confidence (mentions) and NumArticles (events) Automated Full-text retrieval Automated Translation Apply filters on confidence (mention)
Human event validation (45 per cent true positives)
Train Machine Learning Classifier (main IV: dictionary-based detection of protest-words)
Apply Machine Learning Classifier
6,190 EVENTS CLASSIFIED 'PE'
Human Validation (58.5 per cent true positives)
PROTEST EVENT DATASET(N = 3.564)

Figure 1. Process of working with GDELT in protest event analysis.

GDELT's *datafication* phase. Neither is CAMEO the only typology of event classification through automated means (cf., Schrodt & Yonamine, 2013); nor may we necessarily agree with its very definition[9] of protest. In fact, CAMEO's protest event codes do suggest a narrow "repertoire of contention" (Tilly, 1995), i.e. the toolbox of actions established among protestors, with a focus on resource-intensive physical encounters that ignores digital (or other) enrichments of the protest repertoire in recent decades. Thus, when speaking of protest events based on GDELT, we must keep in mind that CAMEO is a typology optimized for machine coding of newspaper reports that prides itself in shifting the focus of event classification from the inter-state to the sub-state level (Schrodt, 2012). This understanding is rooted in the study of peace and conflict rather than collective action and social movements, making the use of data for these objectives less pertinent.

While there are numerous ways to access GDELT, from Google Big Query to SQL, we opted for an approach based on GDELT's *MasterFilesList*, or "the most faithful copy of the incoming data" (The GDELT Project, 2021). We used an automated Python script to iterate over all zipped Event Files on the *MasterFilesList*, to unpack these archives, extract entries with the abovementioned criteria, and store all information on these into a new CSV file. In total, this approach yielded a matrix of 379,747 unique events with 61 variables for each event. GDELT's Event Data is based on events as units of analysis, meaning that every row in that matrix has one unique event ID. However, each of these events can be "mentioned" by several newspaper articles (i.e., stories). This information is stored in a numeric variable labeled "NumArticles." Yet, the Event Database only reveals the number of articles and the *first* URL to a story mentioning the event.

To validate GDELT's information about an event and to perform additional coding tasks (cf., Hutter, 2014), the actual full text of a story is naturally of interest to researchers. This information, however, is invisible. To render this missing (but crucial) information visible for the validation of the protest events, one possibility would be to retrace the single URL to a story associated with an event. Yet, especially when using retrospective data on online newspaper stories over several years, links may not be accessible anymore, leading to dead ends that make it impossible to scrutinize a story and verify the event. The unavailability of a news story's actual text (which GDELT's classification is based upon) is particularly problematic considering the abovementioned problem of false positives. Copyrights and limitations in data storage capacities[10] render the provision of news media texts unfeasible or impossible, which leaves researchers little opportunity to critically engage with the data or to reproduce the analyses that GDELT's event records are based upon. These external restrictions apply to other data repositories like MediaCloud[11] as well.

The steps up to this point revealed a tension between the promise of "an open data firehose" (The GDELT Project, n.d.) and the computational and human resources required to sample a dataset of protest events in each country over a given time. One may argue that neither SQL queries nor writing Python scripts to parse the (at the time of writing) more than 697 thousand zip-archives in GDELT's *MasterFilesList* can be called standard skills taught in most social science departments. Yet, as explained, even possessing such skills does not make critical engagement with the data straightforward. As the purpose of our research went beyond a mere description of the number of events of type x in country y in year z, we needed to engage more critically with the problem of false-positive cases and missing information in many of GDELT's variables. Accordingly, we could not rely solely on a single hyperlink (that may not be functional anymore) to corroborate the information provided by GDELT. Instead, we decided to enrich the *Events* data with *all* stories that GDELT had initially parsed to get information on an event.

In a separate dataset called *Mentions*, GDELT provides links to all stories reporting a single event. Unlike the Events dataset, in which one row constitutes one unique event, the Mentions data is

[9] In fact, the CAMEO codebook even lacks a comprehensive definition of protest.

[10] According to GDELT, one year of their Global Knowledge Graph data amounts to 2.5 Terrabyte of data. In total, GDELT claims to record a quarter-billion events.

[11] https://mediacloud.org/.

ordered around unique mentions. This means both a story's URL and an Event ID can appear multiple times in that dataset, as a single story can mention multiple events, and multiple stories can mention the same event. Again, we accessed the Zip-archives provided through GDELT's MasterFileList and searched for the URLs of stories that mentioned one of the 379,747 events we extracted from the Events dataset. In total, we searched through more than 400,000 CSV files and identified more than 982,000 mentions. These are in addition to the single URL provided in the Events dataset, meaning we had a total of more than 1,361,000 URLs that (according to GDELT) contain information on protest events in the six country cases between 2015 and 2020.

In sum, the invisible processes behind the variables and attributes attached to protest events on GDELT require considerable human and computational labor that is often rendered invisible or not given space in a published article, where the focus is on the analysis and its results. Nevertheless, they are relevant for other researchers as well as for analysts, journalists, and policymakers who draw on these data without the critical reflections of a researcher in the social sciences or the humanities. Leaving that work to computer scientists curating and processing the data shifts our focus to the "data positivism" (Fuchs, 2017) that imbues the data with authority, in no small part due to its size being construed as a proxy for completeness, and the trust given to invisible computational processing performed by the platform. While some of the abovementioned research employing data from GDELT does mention such issues in passing, we also need critical engagement with the processes themselves.

Addressing the false positives problem through machine learning

With the large dataset, a solely manual coding to validate the protest events (PEs, including the variables relevant to the project) returned by GDELT would have been too time-consuming. Our inspections of random data samples revealed a high number of false-positive entries, both in the events data and the mentions data. A false positive meant that an *event* was labeled PE despite it not being an act of *protest*, it happening in a different place, at a different time, or not happening at all. A *mention* may also be coded as reporting on an event, even though an inspection of the actual story would reveal no mention of that event. Being aware of the problem of many false positives meant that we needed a way to validate the protest events. Yet, without the actual texts of the news stories that GDELT's coding is based on, we had no way of verifying the information provided by GDELT. At the same time, the sheer amount of data made human coding unfeasible.

To account for this, we opted for a semi-supervised approach to data classification, aiming to exclude false-positive entries from the Event dataset. We drew a random sample of 1,000 events and let four expert coders evaluate all stories on each of these events. To that end, human coders checked all URLs from the Events and Mentions datasets for each event in the sample. When links were unavailable, coders used the Internet Archive's Wayback Machine (https://archive.org/web/) to obtain historical snapshots of a website (where possible) and used browser-based translation services in case of language barriers.[12] This process can be seen as an approximation to and uncovering of GDELT's inner workings. Retracing some of GDELT's steps illustrates the intermediary role of technology and the agency of platforms: in our case, the question of whether a website is accessible through the Wayback Machine[13] is not in our hands – neither is the quality of translation algorithms required to translate non-English texts as a base for event coding.

Consequently, even though we retraced the way data on GDELT was processed, we needed to do so by employing other platforms (Wayback Machine, translation algorithms) with invisible data processing for which we cannot account ourselves. Nevertheless, this was the only way forward available for us to validate the protest events. As a result of this process, we coded a binary Protest Event variable; *True* for events that could be validated and *False* for events that could not be validated. Accordingly,

[12] For more information on the coding instructions and inter-coder reliability, see the supplement.

[13] Admittedly, the real-time webcrawling of GDELT does not face that problem: at the same time, since its list of sources relies on Google News, the question whether a source is deemed "newsworthy" is made by google, not GDELT.

human coders labeled 448 events as True and 552 as False. Even allowing for a margin of error (e.g., due to non-archived websites[14]), this ratio of false-positive cases in the data supports the above-mentioned criticism that Wang et al. (2016) voiced regarding the actual protest event coverage of GDELT.

Given the high number of false positives and the labor intensity of human coding, we trained a semi-supervised Machine Learning (ML) classifier on the manually-coded data. We experimented with several different techniques, including logistic regression, naïve Bayes, and Support Vector Machines (SVM). We tested a variety of models with different combinations of independent variables from the GDELT Event dataset,[15] none of which yielded sufficient results in terms of precision and recall that would have allowed us to classify the Event dataset into True and False positives confidently. In other words, the machine learning algorithms we tested could not detect a pattern in the GDELT data to identify false-positive cases. The results of the process suggested that false-positive cases are not systematic, allowing for reliable (yet invalid) comparisons across time and place. The results were not acceptable for scientific inquiry, so we opted for investing more time, human capital, and computational power in testing alternative approaches.

Collecting additional data and working with classifiers

In light of the results from the previous step, we determined to enrich the GDELT events data with similar information as that available to the human coders who had validated the sampled events. They relied on each story's text that mentioned an event logged in GDELT's *Event* and *Mentions* data. To automate this process, we used the Python package Newspaper 3k (Ou-Yang, n.d.) to assist us with accessing (where possible) the full text of stories on the web. For the 1,000 events of the ML test and training sets, this amounted to 9,194 queried URLs. Whenever a non-English text was retrieved, we translated it with the Opus MT translator available through the Language Technology Research Group at the University of Helsinki (Tiedemann & Thottingal, 2020). Compared to similar proprietary translators (e.g., Google Translate or DeepL) we found the use of university-developed open models preferable for academic research to for-profit companies' tools, and the considerable costs with accessing their services. Opus MT provides pre-trained neural network translation models for a variety of different languages. These neural networks greatly benefit from GPU (graphical processing unit) instead of CPU power in a Python environment. Therefore, we wrote scripts in Python and executed these on a (paid) remote computing service provided by the University of Copenhagen to simultaneously access multiple computers with high-powered GPU. Doing so, allowed for a reasonably fast batch-wise translation of non-English stories.

To enhance validity, reduce costs and optimize the use of computational power, we further limited the Mentions data to entries with a "confidence" score of at least seven. GDELT's confidence variable in the Mentions dataset indicates the degree to which GDELT's algorithms were sure that a story did mention the event that was coded. It is measured on a ten-point scale,[16] with ten indicating the highest confidence. This step limited the number of stories on the events in our test and training set to 4,776. However, we must bear in mind that the textual data may be noisy, as the collection of website data often yielded errors or retrieved texts that might have changed since the original story was published. Paradoxically, our task of identifying noise (i.e., false positives) brought us to a point where we were forced to introduce even more noise. Ultimately, we attempted to turn this textual data into a useful independent variable for the Machine Learning classifier. To do so, we developed a dictionary of

[14] See p. 22 for detailed information on sources of error.

[15] E.g. the Goldstein scale which measures the disruptiveness of an event, the tonality of reporting on an article, the number of articles, the location, etc.

[16] The variable is actually measured on a scale from 0–100 but it only changes in 10-point increments (i.e. there is no 12 or 18, only a 10 or a 20). Hence, despite of being presented as a 0–100 scale, suggesting a greater level of precision, we refer to it as a 10-point scale.

English words indicating protest events.[17] Converting the input strings (i.e., stories) to lower-case, we could calculate the number of times a word from the protest dictionary could be found in a story.

Since the unit of classification was *Events*, not *Stories/Mentions*, we calculated the following variables on event-level:

- The number of stories for each event that had at least one match with the dictionary;
- The number of total matches with the dictionary in all stories mentioning an event;
- A three-level factor variable that measured whether most stories on an event had at least one match with the dictionary, contained no valuable info (text too short, broken links, etc.) or contained article-texts without matches in the dictionary;
- A binary variable that indicated whether any of the stories on an event contained at least one match with our dictionary.

We added these as independent variables into the ML-classifier, finding that the classification results for the test-set had thus improved significantly. Enriching the GDELT Event data with as much textual information as possible turned out to be the only way of ensuring a reasonable validation of the data. Like in any classification exercise, we had to consider a trade-off between recall and precision, or roughly speaking, the number of false negatives we were willing to accept after classification versus the number of false positives.

As the classifier's objective was to reduce the number of events that would later be manually coded, we opted to emphasize recall over precision. We aimed to limit the number of cases labeled as false by the classifier but which were actual protest events, while at the same time accepting that the classifier labeled events as protest events that were actually not protest events, as the latter would be excluded manually at a later stage in the process. Among the ML-classifiers we tested, we identified a Support Vector Machine with linear Kernel[18] that produced the best results, with a Precision of 0.74 and a Recall of 0.83 for the test-set data.

As shown, to validate the data, we needed to render some of the invisibilities the GDELT database introduces visible. Consequently, we reverse-engineered the classification of events to reconstitute the basis whereupon GDELT labeled protest events in news media texts. That is, we worked with much larger datasets than the protest event data initially provided by GDELT. This also meant that we introduced various steps in the processing of the data (such as classifiers, dictionaries, models, and computational power). While we achieved an acceptable result (and could with a high level of certainty differentiate protest events from false positives), it was the fruit of step-wise decisions, more human and computational labor, and trial and error as we went along.

Large-scale application

To apply the trained classifier to the entire dataset, we reproduced the additional data collection that we conducted on the test and training set for our entire Event dataset. However, we had to impose a cutoff on the amount of data, given the resource intensity of automated data collection and translation. Therefore, we only classified events with at least eight stories mentioning that event, the 60th percentile of the *NumArticles* variable. Imposing the stories-based filter on events and the confidence-based filter on stories left us with more than 146,000 events and 362,279 unique story URLs to query. After the automated retrieval of these stories, more than 68,000 texts turned out to be non-empty, non-duplicate, and non-English, amounting to 180 million characters to be translated by the neural network translation models.

[17] "demonstration", "protest", "rally", "march", "parade", "riot", "strike", "boycott", "sit-in", "crowd", "mass", "picket", "picket line", "blockade", "mob", "flash mob", "revolution", "rebellion", "demonstrations", "protests", "rallies", "marches", "parades", "riots", "strikes", "boycotts", "sit-ins", "crowds", "masses", "pickets", "picket lines", "blockades", "mobs", "flash mobs", "revolutions", "rebellions", "clash", "demonstrate", "campaign", "protester", "protesters".

[18] Formula: "Protest_event ~ IsRootEvent + MonthYear + GoldsteinScale + AvgTone + fulltext_factor_result".

An application of the SVM as outlined above on that dataset led to 91,721 events labeled as Protest Event and 54,756 events labeled as No Protest Event. While we minimized the false negatives among the No Protest Events, we must still address the false positives among the Protest Events. Therefore, we opted for another round of human coding that sought to validate the PE prediction, thus sorting out the true positives from the false positives. Since 91,721 cases are beyond the capacities of five coders (who worked on coding the data), we shifted the NumArticles filter from the 60th percentile to the 95th percentile of events, i.e., we filtered for events that had been mentioned in at least eleven articles.

The underpinning reasoning was well in line with other protest event analyses arguing that the more reports on an event can be found, the more likely it is that it really took place and the more impactful it might have been. As our initial research interest was locating the main protest events in our six country cases, it was safe to assume that these actions would be among those being more repeatedly covered in the media. Imposing this additional penalty to the classified results left 3,863 events labeled No Protest Event and 6,190 events labeled as Protest Event. After securing inter-coder reliability on a random sample of 100 events labeled PE (Fleiss' Kappa = .896), three human coders established the validity of all events that were classified as protest events, using the same coding rules we applied for the test and training sample. In line with the precision score of our classifier, we expected a fair number of false positives that could be filtered out through this additional step. Indeed, human coders validated 3,564 of the 6,190 cases (58.5%) labeled protest events by the classifier.

Ultimately, this process was merely instrumental in pursuit of a dataset of protest events for a set of given places over a specific period of time that can be used for further analysis. Nonetheless, it allowed us to identify some recurring patterns of error. First, in less than three per cent of all events, we were not able to identify a single working, translatable story. As expected, this number rises the further we go back in time, yet never exceeds five per cent. Thus, the risk of broken links and unavailable data when working with historical websites must not be ignored but can be alleviated when using archival tools like the WayBack Machine and selecting events above a certain count threshold for stories. Second, we found that GDELT's mislabeling of PE's had several reasons. While we did not exactly quantify the sources of error, we found that first, many false positives contained no reference to protest at all, and second, the automated geo-classification of GDELT caused mislabels.[19] Only on rare occasions, we found events outside our date-range in the data. We can assume these issues to be rooted in GDELT's location and event classification algorithms, which in turn leads us to question how many instances of protest in our countries we might have missed due to relying on these imposed labels. While little can be done in this regard, except reproducing the massive and imposing work of data collection and processing by GDELT, we must nonetheless be aware of these sources of bias affecting our own and any other findings based on big data news repositories.

Taken together, the description of our process outlined a lot of "trial-and-error" but primarily illustrated the amount of work required to turn GDELT information into a PEA dataset (see, Figure 1). While the analysis of the processed data will be discussed elsewhere, it is noteworthy that many values for variables of interest to protest event researchers (e.g., type of actor initiating the event, actor targeted by the action, exact event location) that GDELT claims to provide are missing. Consequently, we are required not to take GDELT at face value but to manually inspect the news sources, as we have done in our research. In other words, we need to make visible the invisible processes of GDELT to collect, sort, store, and classify the data – or rather those parts that are relevant for our research.

Conclusion

In hindsight, we must contend that GDELT, as one example of a free and open big data repository, is not as free and open as it purports. Neither are the algorithmic inner workings that code GDELT transparent, nor do they yield acceptable results (by most scientific standards). Deep engagement with

[19] This was especially true for the UK and it's former overseas territories, as, for example, many protest events in Hong Kong were labeled "UK".

the Codebook and with the PEA (and other content analyses) literature is necessary to understand and test which variables (like the oblique *confidence* score or the NumArticles variable) and which processes can turn a "greedy" lump of big data into a useful dataset for social research. The idea that advancements in the automation of data collection, storage, and processing provide access to substantial knowledge via a few taps on the touchpad must be called misleading at best. Indeed, even if GDELT may not make promises about such simple access to their data, the fact that major organizations such as the European Union, the US army and major financial corporations use its data in an uncritical manner means that the platform enables an interpretation of it as providing easily accessible data. Instead, our own research process documents the numerous and sometimes invisible decisions taken through the usage of one event typology or the other, one news aggregation service or the other, or of one translation API or the other.

It also documents the various techniques we used, from API programming to webscraping, from human coding to dictionary-based content analyses to machine learning classification – which require human labor, expertise, computational power, time, and economic resources. None of these aspects disqualifies GDELT *per se* – the data accumulated and provided by GDELT are a rich resource, but they tell a cautionary tale: while it is tempting to take big data at face value and ascribe authority to size, scientific diligence requires us to highlight the procedural nature of data by spelling out the various decisions taken consciously or unconsciously, pragmatically or driven by theory, both by the creators and algorithms and by the users of GDELT. In that sense, researchers, journalists, governments, and think tanks may be well-advised not to treat GDELT as a quick-fix for answering substantive questions. As any data, big data news repositories have in-built biases and require rigorous processing before they can be used for valid and reliable analyses. At the same time, our research shows that these data are indeed never raw but always "cooked" (Gitelman, 2013), as they have been created by news media institutions and then processed, sorted, and archived by GDELT – often in invisible and opaque ways. While some of the data processing and the resulting invisibilities were made visible through our research, the underlying epistemological problems of the trust and authority given to big data news repositories remain.

We may legitimately therefore ask why news media data repositories are used by so many researchers, journalists, NGOs, governments, decision-makers, and think tanks when the data quality is so poor? Yet, the alternatives are often resource-intensive, biased, and are not imbued with the same credibility and authority when based on smaller datasets. Moreover, the promises of GDELT and similar platforms are tempting, and the actual cost of the data often remains invisible when presenting the results of such research. As our experience shows, using these data in a way that lives up to expectations of scientific rigor in the social sciences and humanities (and for us in social movement studies, and particularly PEA) is perhaps even more work-intensive, requiring both human and computational labor. More generally, the resource scarcity making civil society dependent on such data repositories may be true of global inequalities in data access which the low-quality visible data in GDELT could further compound, hampering resource-poor researchers and practitioners seeking to conduct sound, reliable and transparent analyses.

While we have outlined a way forward with using protest event data from GDELT for PEA, this article should not be used uncritically as a research manual. Instead, it should be treated with the same caution as any context-specific research. Nonetheless, we believe that the documentation of our process and the critical discussion of steps taken along that way provide a valuable resource to be adopted and amended by other researchers (see https://github.com/walfaelschung/GDELT_flow). Through this, we made the underlying processes of working with such data visible while at the same time highlighting that such an approach is based on researchers' specific sets of expertise and distinct materialities, and is dependent on access to computational power and methods. The critical discussion of this piece is thus supplemented by an open-access step-by-step process documentation, including code snippets, that illustrates the critical junctures of our work and can inform decisions in similar, future, projects.

In our turn, we made visible the bearing that the underlying processing of the data by GDELT has on the quality of data and its validation. The invisibilities we uncovered render the reproduction of our, by contrast transparent, approach difficult, if not impossible. As such, while we may have some understanding of the bias we introduce by using specific sources (usually newspapers) with more traditional approaches in PEA, we cannot understand the invisibilities (and bias) introduced by big data news media repositories. The logic of scientific production that often follows (particularly in the age of big data) is one where results must come quick (C. Neumayer & Rossi, 2016), and it is difficult to secure funding for human coding when data like GDELT is promised to be "available for free." Yet, we need to use such platforms with caution and make visible their underlying processes as relevant for our research to critically engage with and make meaningful, reliable use of the data.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors acknowledge receipt of funding for this research from the Volkswagen Foundation (grant no. 98144).

Notes on contributors

Matthias Hoffmann is a postdoctoral researcher at the Department of Communication, University of Copenhagen, working on protest, social movements, and social media.

Felipe G. Santos is a Postdoctoral Research Fellow at the Department of Sociology and Criminology of the City, University of London. He is interested in how different social groups and political ideologies push for long-lasting societal changes. With this aim, his research covers social movements, political parties, and other civil organizations, spanning from the left to the far right.

Christina Neumayer is associate professor of media studies at the Department of Communication at the University of Copenhagen.

Dan Mercea is Reader in Digital and Social Change at City, University of London. He has a lasting interest in social movements and the implications flowing from the adoption and repurposing of Internet technologies in various domains of social and political activity.

ORCID

Matthias Hoffmann http://orcid.org/0000-0001-6480-3679
Felipe G. Santos http://orcid.org/0000-0001-7006-2088
Christina Neumayer http://orcid.org/0000-0003-0450-2983
Dan Mercea http://orcid.org/0000-0003-3762-2404

References

Austin Holmes, A., & Baoumi, H. (2016, January 29). Egypt's protests by the numbers. *Carnegie Endowment for International Peace*. https://carnegieendowment.org/sada/?fa=62627

Banks, A. S. (1997). *Cross-nationaltime-series data archive* [English]. https://www.worldcat.org/title/cross-national-time-series-data-archive/oclc/768447979

Bekker, M. (2022). Better, faster, stronger: Using machine learning to analyse south African police-recorded protest data. *South African Review of Sociology*, *52*(1), 4–23. https://doi.org/10.1080/21528586.2021.1982762

Best, R. H., Carpino, C., & Crescenzi, M. J. C. (2013). An analysis of the TABARI coding system. *Conflict Management and Peace Science*, *30*(4), 335–348. https://doi.org/10.1177/0738894213491176

Bolivar, F., Camara, N., Davila Egas, T., Orkun Isa, B., Posadas, C., Rodrigo, T., & Vazquez, S. (2021). Understanding the sustainability framework using Big Data. *BBVA Research*. https://www.bbvaresearch.com/en/publicaciones/global-understanding-the-sustainability-framework-using-big-data/

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bruns, A., Harrington, S., & Hurcombe, E. (2021). Coronavirus conspiracy theories: Tracing misinformation trajectories from the fringes to the mainstream. In M. Lewis, E. Govender, & K. Holland (Eds.), *Communicating COVID-19: interdisciplinary perspectives* (pp. 229–250). Springer International Publishing. https://doi.org/10.1007/978-3-030-79735-5

Bruns, A., Hurcombe, E., & Harrington, S. (2021). Covering conspiracy: Approaches to reporting the COVID/5G conspiracy theory. *Digital Journalism*, 1–22. https://doi.org/10.1080/21670811.2021.1968921

Campante, F., & Yanagizawa-Drott, D. (2018). Long-range growth: Economic development in the global network of air links. *The Quarterly Journal of Economics*, *133*(3), 1395–1458. https://doi.org/10.1093/qje/qjx050

Carvalho, T. (2022). *Contesting austerity: Social movements and the left in Portugal and Spain (2008-2015)*. Amsterdam University Press.

Chalabi, M. (2014, May 6). Kidnapping of girls in Nigeria is part of a worsening problem (Updated). *FiveThirtyEight*. https://fivethirtyeight.com/features/nigeria-kidnapping/

Christensen, D. (2019). Concession stands: How mining investments incite protest in Africa. *International Organization*, *73*(1), 65–101. https://doi.org/10.1017/S0020818318000413

Christensen, D., & Garfias, F. (2018). Can you hear me now? How communication technology affects protest and repression. *Quarterly Journal of Political Science*, *13*(1), 89–117. https://doi.org/10.1561/100.00016129

Claassen, C., & Gibson, J. L. (2016). *Macro-tolerance and protest: Does a culture of political intolerance dampen dissent?*

Consoli, S., Pezzoli, L. T., & Tosetti, E. (2021). Emotions in macroeconomic news and their impact on the European bond market. *Journal of International Money and Finance*, *118*, 102472. https://doi.org/10.1016/j.jimonfin.2021.102472

David Williams, O., Yung, K. C., & Grépin, K. A. (2021). The failure of private health services: COVID-19 induced crises in low- and middle-income country (LMIC) health systems. *Global Public Health*, *16*(8–9), 1320–1333. https://doi.org/10.1080/17441692.2021.1874470

Dearing, J. W., Rogers, E. M., & Rogers, E. (1996). *Agenda-Setting*. SAGE.

Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society*, *2*(2), 2053951715617185. https://doi.org/10.1177/2053951715617185

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Dos Santos, R. F., Perkins, T. K., Wood, C. D., Meyer, W. D., Garfinkle, N. W., Enscore, S. I., Wang, X., Selig, L. A., & Calfas, G. W. (2017). *Social.* and Political Event Data to Support Army Requirements (ERDC/CERL TR-17-40; Military Facilities Engineering Technology). U.S. Army Engineer Research Development Center.

Drakos, K., & Gofas, A. (2006). The devil you know but are afraid to face: Underreporting bias and its distorting effects on the study of terrorism. *Journal of Conflict Resolution*, *50*(5), 714–735. https://doi.org/10.1177/0022002706291051

Earl, J., Martin, A., McCarthy, J. D., & Soule, S. A. (2004). The use of newspaper data in the study of collective action. *Annual Review of Sociology*, *30*(1), 65–80. https://doi.org/10.1146/annurev.soc.30.012703.110603

Earl, J., Soule, S. A., & McCarthy, J. D. (2003). Protest under fire? Explaining the policing of protest. *American Sociological Review*, *68*(4), 581–606. https://doi.org/10.2307/1519740

Eck, K. (2012). In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Cooperation and Conflict*, *47*(1), 124–141. https://doi.org/10.1177/0010836711434463

The Economist. (2020, March 10). *Political protests have become more widespread and more frequent. The Economist*. https://www.economist.com/graphic-detail/2020/03/10/political-protests-have-become-more-widespread-and-more-frequent

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Fengcai, Q., Jinsheng, D., & Li, W. (2020). An online framework for temporal social unrest event prediction using news stream. *2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 176–182. https://doi.org/10.1109/CyberC49757.2020.00036

Ferreira, L. N., Hong, I., Rutherford, A., & Cebrian, M. (2021). The small-world network of global protests. *Scientific Reports*, *11*(1), 19215. https://doi.org/10.1038/s41598-021-98628-y

Fuchs, C. (2017). From digital positivism and administrative big data analytics towards critical digital and social media research! *European Journal of Communication*, *32*(1), 37–49. https://doi.org/10.1177/0267323116682804

Fu, K., & Zhu, Y. (2020). Did the world overlook the media's early warning of COVID-19? *Journal of Risk Research*, *23*(7–8), 1047–1051. https://doi.org/10.1080/13669877.2020.1756380

Galambos, L. (1975). *The public image of big business in America, 1880-1940: A quantitative study in social change*. Johns Hopkins University Press.

The GDELT Project. (2015, February 19). *GDELT 2.0: Our global world in realtime*. https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

The GDELT Project. (2021, January 22). *A behind-the-scenes look at how we think about master file formats and timestamping*. https://blog.gdeltproject.org/a-behind-the-scenes-look-at-how-we-think-about-master-file-formats-and-timestamping/

The GDELT Project. (n.d.). *The GDELT project: Watching our world unfold*. Retrieved November 18, 2021, from https://www.gdeltproject.org

Gitelman, L. (2013). *Raw data is an oxymoron*. MIT Press.

Guo, L., & Vargo, C. (2020). "Fake News" and emerging online media ecosystem: an integrated intermedia agenda-setting analysis of the 2016 U.S. Presidential election. *Communication Research*, *47*(2), 178–200. https://doi.org/10.1177/0093650218777177

Haig, C. S., Schmidt, K., & Brannen, S. (2020, March 2). The Age of Mass Protests: Understanding an Escalating Global Trend. Center for Strategic and International Studies. https://www.csis.org/analysis/age-mass-protests-understanding-escalating-global-trend

Halkia, M., Ferri, S., Papazoglou, M., Van Damme, M.-S., & Thomakos, D. (2020). Conflict event modelling: Research experiment and event data limitations. *Proceedings of AESPEN 2020*, 42–48. https://aclanthology.org/2020.aespen-1.8/

Hammond, J., & Weidmann, N. B. (2014). Using machine-coded event data for the micro-level study of political violence. *Research & Politics*, *1*(2). https://doi.org/10.1177/20531680145399

Hopp, F. R., Fisher, J. T., & Weber, R. (2020). Dynamic transactions between news frames and sociopolitical events: An integrative, hidden markov model approach. *Journal of Communication*, *70*(3), 335–355. https://doi.org/10.1093/joc/jqaa015

Hopp, F. R., Schaffer, J., Fisher, J. T., & Weber, R. (2019). iCoRe: The GDELT interface for the advancement of communication research. *Computational Communication Research*, *1*(1), 13–44. https://doi.org/10.5117/CCR2019.1.002.HOPP

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), 8. https://doi.org/10.1111/lnc3.12432

Hutter, S. (2014). Protest event analysis and its offspring. In D. Della Porta (Ed.), *Methodological practices in social movement research*. Oxford Scholarship Online.

Jäger, K. (2018). The limits of studying networks via event data: Evidence from the ICEWS dataset. *Journal of Global Security Studies*, *3*(4), 498–511. https://doi.org/10.1093/jogss/ogy015

Jenkins, J. C., & Perrow, C. (1977). Insurgency of the powerless: Farm worker movements (1946-1972). *American Sociological Review*, *42*(2), 249–268. https://doi.org/10.2307/2094604

Katzenbach, C. (2021). "AI will fix this" – The technical, discursive, and political turn to AI in governing communication. *Big Data & Society*, *8*(2), 20539517211046184. https://doi.org/10.1177/20539517211046182

Kolanovic, M., & Krishnamachari, R. T. (2017). *Big data and AI strategies: Machine learning and alternative data approach to investing*. J.P. Morgan.

Kriesi, H., Wüest, B., Lorenzini, J., Makarov, P., Enggist, M., Rothenhäusler, K., Kurer, T., Häusermann, S., Wangen, P., Altiparmakis, A., Borbáth, E., Bremer, B., Gessler, T., Hunger, S., Hutter, S., Schulte-Cloos, J., & Wang, C. (2020). *PolDem-protest dataset 30 European countries, Version 1*. https://poldem.eui.eu/downloads/pea/poldem-protest_30_codebook.pdf

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (Fourth ed.). SAGE.

Kurer, T., Häusermann, S., Wüest, B., & Enggist, M. (2019). Economic grievances and political protest. *European Journal of Political Research*, *58*(3), 866–892. https://doi.org/10.1111/1475-6765.12318

Kwak, H., & An, J. (2014). A first look at global news coverage of disasters by using the GDELT dataset. In L. M. Aiello & D. McFarland (Eds.), *Social informatics: 6th International Conference, Socinfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings* (1st edition). Springer.

LaFree, G. (2010). The Global Terrorism Database (GTD): Accomplishments and challenges. *Perspectives on Terrorism*, *4*(1), 24–46.

Lakoff, G. (1990). *Don't think of an elephant: Know your values and frame the debate*. Chelsea Green Publishing Co.

Leetaru, K. (2014, May 30). Did the Arab spring really spark a wave of global protests? *Foreign Policy*. https://foreignpolicy.com/2014/05/30/did-the-arab-spring-really-spark-a-wave-of-global-protests/

Leetaru, K., & Schrodt, P. A. (2013). *GDELT: Global data on events, location and tone, 1979-2012*.

Levin, N., Ali, S., & Crandall, D. (2018). Utilizing remote sensing and big data to quantify conflict intensity: The Arab spring as a case study. *Applied Geography*, *94*, 1–17. https://doi.org/10.1016/j.apgeog.2018.03.001

Malik, M., Hopp, F. R., Chen, Y., & Weber, R. (2021). Does regional variation in pathogen prevalence predict the moralization of language in COVID-19 news? *Journal of Language and Social Psychology*, *40*(5–6), 653–676. https://doi.org/10.1177/0261927X211044194

Manacorda, M., & Tesei, A. (2020). Liberation technology: Mobile phones and political mobilization in Africa. *Econometrica*, *88*(2), 533–567. https://doi.org/10.3982/ECTA14392

Mattoni, A., & Pavan, E. (2018). Politics, participation and big data. Introductory reflections on the ontological, epistemological, and methodological aspects of a complex relationship [Data set]. *Partecipazione & Conflitto*, *11*(2), 313–331. https://doi.org/10.1285/I20356609V11I2P313

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McAdam, D. (19821999). *Political process and the development of black insurgency, 1930–1970*. University of Chicago Press.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*(2), 176–187. https://doi.org/10.1086/267990

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Metternich, N. W., Dorff, C., Gallop, M., Weschle, S., & Ward, M. D. (2013). Antigovernment networks in civil conflicts: How network structures affect conflictual behavior. *American Journal of Political Science*, *57*(4), 892–911. https://doi.org/10.1111/ajps.12039

Neumayer, C.(2022). Content, form and reception: Perspectives from digital media data. In P. Vossen & A. Fokkens (Eds.), *The perspective web* (pp. 143–155). Cambridge University Press.

Neumayer, M. M., & Poell, T. (2019). Introduction. In C. Neumayer, M. Mortensen, & T. Poell (Eds.), *Social media materialities and protest: Critical reflections* (pp. 1–14). Routledge.

Neumayer, C., & Rossi, L. (2016). 15 years of protest and media technologies scholarship: A sociotechnical timeline. *Social Media + Society*, *2*(3), 2056305116662180. https://doi.org/10.1177/2056305116662180

Neumayer, C., Rossi, L., & Struthers, D. M. (2021). Invisible data: A framework for understanding visibility processes in social media data. *Social Media + Society*, *7*(1), 2056305120984472. https://doi.org/10.1177/2056305120984472

Odziemkowska, K., & Henisz, W. J. (2021). Webs of influence: Secondary stakeholder actions and cross-national corporate social performance. *Organization Science*, *32*(1), 233–255. https://doi.org/10.1287/orsc.2020.1380

Olzak, S. (1992). *The dynamics of ethnic Competition and Conflict*. Stanford University Press.

Ortiz, A., & Rodrigo, T. (2018). Monitoring global trade support in real time using Big Data. *BBVA Research*. https://www.bbvaresearch.com/wp-content/uploads/2018/07/Exploring-the-global-trade-and-protectionism-in-real-time-using-Big-Data_.pdf

Ou-Yang, L. (n.d.). Newspaper. *Github*. https://github.com/codelucas/newspaper

Ponticelli, J., & Voth, H.-J. (2020). Austerity and anarchy: Budget cuts and social unrest in Europe, 1919–2008. *Journal of Comparative Economics*, *48*(1), 1–19. https://doi.org/10.1016/j.jce.2019.09.007

Portos, M. (2021). Grievances and public protests: Political mobilisation in Spain in the age of austerity. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-53405-9

Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: An armed conflict location and event dataset: Special data feature. *Journal of Peace Research*, *47*(5), 651–660. https://doi.org/10.1177/0022343310378914

Scheufele, D. A. (1999). Framing as a Theory of Media Effects. *Journal of Communication*, *49*(1), 103–122. https://doi.org/10.1111/j.1460-2466.1999.tb02784.x

Schrodt, P. A. (2012). *CAMEO conflict and mediation event observations event and actor codebook*. http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

Schrodt, P. A., & Yonamine, J. E. (2013). A guide to event data: Past, present, and future. *All Azimuth: A Journal of Foreign Policy and Peace*, *2*(2), 5–22.

Sundberg, R., & Melander, E. (2013). Introducing the UCDP georeferenced event dataset. *Journal of Peace Research*, *50*(4), 523–532. https://doi.org/10.1177/0022343313484347

Tiedemann, J., & Thottingal, S. (2020). *OPUS-MT – Building open translation services for the world*. https://helda.helsinki.fi/handle/10138/327852

Tilly, C. (1995). *Popular contention in Great Britain, 1758-1834*. Harvard University Press.

Vargo, C. J., & Guo, L. (2017). Networks, big data, and intermedia agenda setting: An analysis of traditional, partisan, and emerging online U.S. news. *Journalism & Mass Communication Quarterly*, *94*(4), 1031–1055. https://doi.org/10.1177/1077699016679976

Wang, W., Kennedy, R., Lazer, D., & Ramakrishnan, N. (2016). Growing pains for global monitoring of societal events. *Science*, *353*(6307), 1502–1503. https://doi.org/10.1126/science.aaf6758

Wang, D. J., & Soule, S. A. (2012). Social movement organizational collaboration: Networks of learning and the diffusion of protest tactics, 1960–1995. *American Journal of Sociology*, *117*(6), 1674–1722. https://doi.org/10.1086/664685

Ward, M. D., Berger, A., Cutler, J., Matthew, D., Cassy, D., & Ben, R. (2013). *Comparing GDELT and ICEWS event data*.

Welbers, K., Van Atteveldt, W., Bajjalieh, J., Shalmon, D., Joshi, P. V., Althaus, S., Chan, C.-H., Wessler, H., & Jungblut, M. (2022). Linking event archives to news: A computational method for analyzing the gatekeeping process. *Communication Methods and Measures*, *16*(1), 59–78. https://doi.org/10.1080/19312458.2021.1953455

Williams, S. (2020). *Exploration of the Global Database of Events, Language and Tone (GDELT), with specific application to disaster reporting*. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/explorationoftheglobaldatabaseofeventslanguageandtonegdeltwithspecificapplicationtodisasterreporting#strengths-and-limitations

Wright, J., Lennox, R., & Verissimo, D. (2020). *Online monitoring of global attitudes towards wildlife*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3658481

Wu, C., & Gerber, M. S. (2018). Forecasting civil unrest using social media and protest participation theory. *IEEE Transactions on Computational Social Systems*, *5*(1), 82–94. https://doi.org/10.1109/TCSS.2017.2763128

Yesilbas, V., Padilla, J. J., & Frydenlund, E. (2021). An analysis of global news coverage of refugees using a big data Approach. In R. Thomson, M. N. Hussain, C. Dancy, & A. Pyke (Eds.), *Social, Cultural, and Behavioral Modeling: 14th International Conference, SBP-BRiMS 2021, Virtual Event, July 6–9,2021, Proceedings* (Vol.12720, pp. 111–120). Springer International Publishing. https://doi.org/10.1007/978-3-030-80387-2

Yuen, S., Cheng, E. W., Or, N. H. K., Grépin, K. A., Fu, K.-W., Yung, K.-C., & Yue, R. P. H. (2021). A tale of two city-states: A comparison of the state-led vs civil society-led responses to COVID-19 in Singapore and Hong Kong. *Global Public Health*, *16*(8–9), 1283–1303. https://doi.org/10.1080/17441692.2021.1877769

Zheng, C. (2020). Comparisons of the city brand influence of global cities: word-embedding based semantic mining and clustering analysis on the big data of gdelt global news knowledge graph. *Sustainability*, *12*(16), 6294. https://doi.org/10.3390/su12166294