Supplementary Material

"Public debate in the media matters: evidence from the European refugee crisis"

May 7, 2020

Caleb M. Koch[a], Izabela Moise[b], Dirk Helbing[a], Karsten Donnay[c,d,*]

[a]*Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland*
[b]*Swiss Data Science Center, ETH Zurich, Zurich, Switzerland*
[c]*Department of Political Science, University of Zurich, Zurich, Switzerland*
[d]*Department of Politics and Public Administration, University of Konstanz, Konstanz, Germany*

A. Data

A.1. GDELT

GDELT[1] (Global Data on Events, Location and Sentiment) is a platform that extracts, analyses, and reports measures of news articles published online. The recent version of GDELT (used in our paper) improves upon several coding schemes traditionally used in political science empirical work, including an upgraded version of CAMEO coding. Other features, such as location and sentiment analyses, make this dataset appropriate for the purposes of our paper. In this section, we detail the key features of GDELT that pertain to our analyses; more specifics can be found in Leetaru and Schrodt (2013).

It is important to note that the GDELT dataset consists of two versions: the GDELT Knowledge Graph and the GDELT Event Database. The first consists of daily files with more complex algorithms such as article polarity (i.e. in-article sentiment variance). While the added complexity of the former is appealing, the main problem is that the GDELT Knowledge Graph is only available starting in 2013. The GDELT Event Database, however, is available from 2002–*present* but only performs the standard sentiment algorithm. The former may prove useful for scholars aiming at a more nuance analysis of news media rather than 'public-wide debate', which is the focus in the main text.

*Corresponding author, email: donnay@ipz.uzh.ch

[1]`http://gdeltproject.org`

We briefly mention some legal concerns that often appear when using the GDELT dataset. A recent article in *Science* that uses GDELT (Wang et al., 2016) describes the issue as follows:

> *"... while the legal issues surrounding GDELT are opaque (the request of one author for clarification was met with an ambiguous response of "it's a touchy subject"), it appears that one of the main developers of GDELT may have utilized copyrighted resources purchased by the University of Illinois' Cline Center for developing the SPEED dataset. It should be noted that we only utilize the aggregate historical data for comparing GDELT with other event data projects, and only utilize the publicly available articles for our validity analysis. In no way have we accessed any copyrighted material purchased by either the Cline Center or GDELT.* ***As far as we are aware there is no ongoing litigation surrounding the use of GDELT and the data has since been re-established on the web.*** *It has also been utilized by government agencies, Google, and continues to produce reports for Foreign Policy and other publications...,"* (Wang et al., 2016, *SI*, p. 3).

We take the same strategy as above: we also only use publicly available data, and in no way have we accessed any copyrighted material purchased by either the Cline Center or GDELT. This is the same approach used by the several studies that utilize the GDELT dataset, e.g., Hammond and Weidmann (2014), De Juan and Bank (2015), Steinert-Threlkeld et al. (2015), Davis et al. (2017), and Steinert-Threlkeld (2017).

A.1.1. Text sources

As mentioned in the main document, the news text sources for GDELT came from online sources, which includes all international news coverage of AfricaNews, Agence France Presse, Associated Press Online, Associated Press Worldstream, BBC Monitoring, Christian Science Monitor, Facts on File, Foreign Broadcast Information Service, United Press International, Washington Post, and all national and international news coverage of New York Times, Associated Press, and Google News. This is only a subset of the sources from which GDELT draws. A more comprehensive list can be found in Leetaru and Schrodt (2013).

A.1.2. Actor coding and location algorithm

GDELT uses an enhanced TABARI coding system to establish actor assignments, i.e. deciphering the key individuals, countries, and/or politicians discussed in the article. This actor deciphering algorithm works in two steps. First, actors in the article are cross-referenced with an expanded *.agents* dictionary with around 60,000 entries, which also leverages the *WordNet* dictionary. For actors not in the dictionary, actor codes are automatically assembled from a 'named actor' in the article and a 'generic agent' existing in the dictionary, where the dictionary for generic agents has approximately 1,500 entries. Hence, the GDELT algorithm does not discard articles with new actors.

Second, this cross-referencing process goes through several iterations in order to improve the accuracy of associating actors with locations. More specifically, cross-referencing occurs on the (i) raw text, (ii) text where actors are replaced with known countries, (iii) cities replaced with countries, and (iv) text where both actors and cities are replaced. To showcase actor-replacement iteration step, consider the following sentence (provided by GDELT documentation):

> *"Egyptian Minister of Foreign Affairs Mohamed Orabi attended the summit yesterday. While he was there, Orabi pledged support for..."*

During the actor-replacement, the second sentence is replaced with *"While he was there, Egypt pledged support for..."*. This process improves the matching process of diplomatic events where participating political leaders are not important enough to have his / her own entry in TABARI. This multiple iteration step essentially helps cope with complex texts and was found to improve actor coding significantly (see Leetaru and Schrodt, 2013, for details).

If the algorithm does not detect a main actor, then the algorithm will also leave the georeference field of the actor blank. Furthermore, the algorithm assigns actors to countries only if such contextual information is available within the text and otherwise leaves the category blank. Pertinent to our main document, the improved matching process is important for determining, e.g., the 'United Kingdom's' treatment of refugees when the article refers to UK representatives.

In addition to actor coding, articles undergo a full-text geocoding process that identifies the geo-location of every actor.[2] The algorithm first identifies all geo-references in the article; then, each identified actor is assigned to the nearest geo-reference. While this may

[2]Specific details are in Leetaru (2012).

seem naïve, the reported accuracy of the algorithm is quite high (see Leetaru (2012) and Leetaru and Schrodt (2013) for details). One example provided in the documentation is as follows: if the President of Russia attends a meeting in Washington D.C., then the algorithm assigns actor 1's country as Russia and not Washington D.C. If no country reference is identified in the article and/or if no cross-reference is available, then the geo-reference to the actor is left unspecified.

A.2. Asylum data

Note that, for replication purposes, one can download the data from the website or go to our "replication files" folder that accompanies this paper, where we have already downloaded the relevant data files.

- *Total positive decisions* – this dataset can be accessed from the EU database directly: `https://ec.europa.eu/eurostat/databrowser/view/tps00192/default/table?lang=en`. Note that, when we accessed this database in January 2017, data was available quarterly. The database, however, currently aggregates data on an annual basis. As such, for replication purposes, please either (i) request quarterly data from the EU directly or (ii) use the data in our "replication folder" that accompanies this manuscript.

- *Rejected* – this dataset can be accessed from the EU database directly: `https://ec.europa.eu/eurostat/databrowser/view/tps00192/default/table?lang=en`. Similar to above, this database was avaiablle at a quarterly-level when we accessed it in January 2017. The database, however, currently aggregates data on an annual basis. As such, for replication purposes, please either (i) request quarterly data from the EU directly or (ii) use the data in our "replication folder" that accompanies this manuscript.

- *First-time asylum applications* – this dataset can be accessed from the EU database directly: `https://ec.europa.eu/eurostat/databrowser/view/tps00189/default/table?lang=en`. For the purposes of our paper, we aggregated this monthly data to the quarterly level.

A.3. Control variables

All control variables we use come from the EU official database and the OECD database.[3] Below we list each control variable and relevant details so that other scholars may reproduce our results. Note that, for replication purposes, one can download the data from the website or go to our "replication files" folder that accompanies this paper, where we have already downloaded the relevant data files.

- *Unemployment rate* – this dataset can be accessed from the EU database directly: `https://ec.europa.eu/eurostat/databrowser/view/tipsun30/default/table?lang=en`. Unemployment here includes individuals between 15 and 74 who are viable candidates for work, actively seek, yet not able to acquire a job. The metric is seasonally adjusted, in the sense that not as many jobs are offered in the Spring as compared to the Winter holiday season.

- *GDP* – this dataset can be accessed from the EU database directly: `https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=naidq_10_gdp&lang=en`. We use quarterly reported data in units of million-Euros, and we use seasonally and calendarly unadjusted data.

- *Government consolidated gross debt* – this dataset can be accessed from the EU database directly: `https://ec.europa.eu/eurostat/databrowser/view/tipsgo20/default/table?lang=en`. This data represents the gross debt by the general government for each quarter. For the purposes of our study, we use the raw metric (i.e. not adjusted according to GDP) in units of million-Euros.

- *Consumer price index (CPI)* – This dataset can be accessed from the OECD database directly: `https://stats.oecd.org/Index.aspx?QueryId=68144`. The CPI is the weighted average of prices of a basket of consumer goods and services. Data is chain-linked normalized to 2010.

We acquired all data in January 2017.

[3] Only the consumer price index comes from the OECD.

References

C. L. Davis, A. Fuchs, and K. Johnson. State control and the effects of foreign relations on bilateral trade. *Journal of Conflict Resolution*, pages 1–34, 2017.

A. De Juan and A. Bank. The Ba ‘athist blackout? Selective goods provision and political violence in the Syrian civil war. *Journal of Peace Research*, 52(1):91–104, 2015.

J. Hammond and N. B. Weidmann. Using machine-coded event data for the micro-level study of political violence. *Research & Politics*, 1(2):1–8, 2014.

K. Leetaru and P. A. Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. *ISA Annual Convention*, 2(4), 2013.

K. H. Leetaru. Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched Wikipedia. *D-Lib Magazine*, 18(9):5, 2012.

Z. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Flowler. Online social networks and offline protest. *EPJ Data Science*, 4(19):1–9, 2015.

Z. C. Steinert-Threlkeld. Spontaneous collective action: Peripheral mobilization during the arab spring. *American Political Science Review*, 111(2):379–403, 2017.

W. Wang, R. Kennedy, D. Lazer, and N. Ramakrishnan. Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1503, 2016.