GLOBE GLOBAL GOVERNANCE
AND THE EUROPEAN UNION
Future Trends and Scenarios

REPORT

The Empirical Use of GDELT Big Data in Academic Research

Project:	GLOBE – The European Union and the Future of Global Governance
GA:	822654
Call:	H2020-SC6-GOVERNANCE-2018
Funding Scheme:	Collaboration Project

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

DISCLAIMER

This project has received funding from the European Union's Horizon 2020 Research & Innovation programme under Grant Agreement no. 822654. The information in this deliverable reflects only the authors' views and the European Union is not liable for any use that may be made of the information contained therein.

Authors: Angel Saz-Carranza, Pau Maturana and Xavier Quer

CONTENTS

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

ABSTRACT

The use of big data in social sciences research has been a growing trend for a number of years now. In this paper we review the use of the Global Database of Events, Language and Tone -GDELT- project, which registers world events and codes the sentiment associated with their coverage. Supported by Google Cloud's cutting-edge algorithm-based technology, the initiative is able to automatically gather and process data in an unprecedented scale, largely surpassing any prior comparable endeavour. Since its early presentation in 2011, various authors have explored the usability of GDELT and sentiment analysis for academic research purposes in the social sciences. We explore the use of the GDELT databases and identify potentialities and current limitations of GDELT by reviewing the existing academic studies which have used it.

Key words: GDELT, sentiment analysis, big data, tone, text analysis, coding.

INTRODUCTION

The use of big data in social sciences research has been a growing trend for a number of years now (1). Defined as "extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions"[1] (2), its presence in research has increased steadily, as authors uncover an ever-expanding array of applications for it, mostly complementing and sometimes improving traditional empirical methods, from case studies to mainstream econometrics (3).

Under this broad definition, several types of big data can be distinguished. One of these, the focus of this article, is sentiment/semantic big data, mainly characterized for representing the underlying sentiment in large corpora of texts, coded using algorithmic processes (4). Probably, the most ambitious sentiment data base nowadays is the Global Database of Events, Language and Tone, popularly known as GDELT, which in 2018 had over 3.2 trillion datapoints, covering the evolution of global human society throughout the past 200 years (6).

GDELT has unsurprisingly caught the attention of numerous researchers in the social sciences for its vast, likely incomparable amount of data. In this article we first explain what GDELT is and how it works to then assess how it has been used in academic research (6), which applications has it been used for, in which domains, and finally, what are the authors' thoughts on its reliability and methodological robustness (7).

SENTIMENT ANALYSIS

According to Feldman (2013; page 83):[2]

"Sentiment analysis (or *opinion mining*) is defined as the task of finding the opinions of authors about specific entities." The general architecture of a generic sentiment analysis system is the following. "The input to the system is a corpus of documents in any format (PDF, HTML, XML, Word, among others). The documents in this corpus are converted to text and are pre-processed using a variety of linguistic tools such as stemming, tokenization, part of speech tagging, entity extraction, and relation extraction. The system may also utilize a set of lexicons and linguistic resources. The main component of the system is the document analysis module, which utilizes the linguistic resources to annotate the pre-processed documents with sentiment annotations. The annotations may be attached to whole documents (for document-

[1] Oxford Dictionary (2020).

[2] Feldman, R. (2013). Techniques and Applications of Sentiment Analysis. *Communications of the ACM* (April 1, 2013)

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

based sentiment), to individual sentences (for sentence-based sentiment) or to specific aspects of entities (for aspect-based sentiment). These annotations are the output of the system and they may be presented to the user using a variety of visualization tools."

The graph below illustrates the aforementioned process.

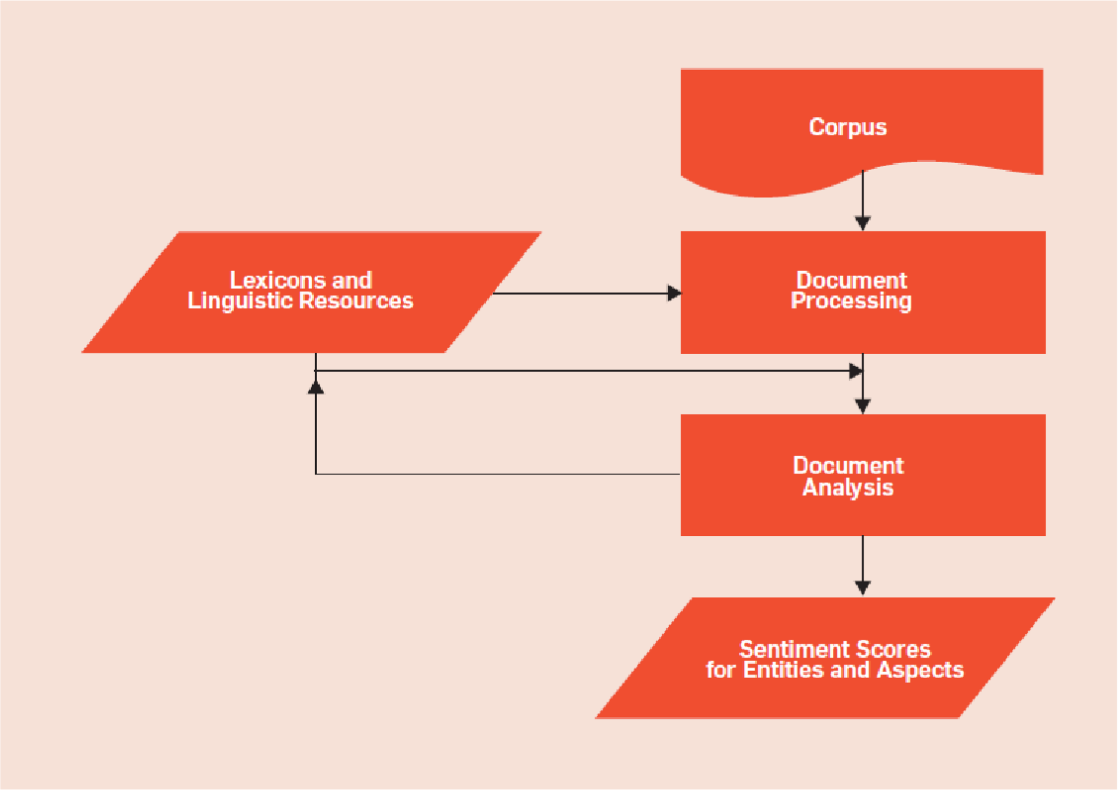
Corpus
Lexicons and Linguistic Resources
Document Processing
Document Analysis
Sentiment Scores for Entities and Aspects

Figure 1. Architecture of a generic sentiment analysis system. Retrieved from Feldman (2013).

THE GLOBAL DATABASE OF EVENTS, LANGUAGE AND TONE (GDELT)

The Global Database of Events, Language, and Tone[3] is a project created by Kalev Leetaru and Georgetown University, self-described as an initiative aimed at leveraging data "to construct a catalog of human societal-scale behavior and beliefs across all countries in the world, connecting every person, location, count, theme, news source , and event across the planet into a single massive network that captures what's happening around the world, what's its context and who's involved, and how the world

[3] GDELT.

Funded by the Horizon 2020 Framework Programme of the European Union. Grant agreement number 822654

is feeling about it, every single day." [4] In other words, the project's goal is to register and code as many world events as possible and the sentiment associated with their coverage.

GDELT does so by collecting all online news published every 15 minutes, machine-translating them into English, and coding each news article by theme, sentiment and tone, location, and entity (organizations and persons), among others. Through cutting-edge algorithm-based technology supported by Google Cloud, GDELT's gathering and data processing capabilities largely surpass any prior comparable endeavor, generating over 1 trillion data points per year.

In particular, GDELT generates several databases, two of which are of relevance for this assessment. Firstly, GDELT Events is a database of events concerning cooperation and conflict relationships between two or more entities which uses CAMEO (Conflict and Mediation Event Observations), a framework for coding event data typically used for events that merit news coverage and generally applied to the study of political news and violence. Secondly, GDELT Global Knowledge Graph is a database of news articles where entities, themes, locations and tone are coded. The figure below compares the two databases.

GDELT Events	GDELT GKG
All news produced by news sources machine-translated into English	
• Code event (if it exists) in news articles using CAMEO: ○ Identify entities, ○ Relationship based on the Goldstein Cooperation-Confrontation scale; • Code themes, • Code locations, • Aggregate news into events.	• Code for ○ tone, ○ themes, ○ entities in each news article, ○ locations.
Database of events going back to 1979	Database of news articles going back to 2014

The GKG database in detail

To date, GDELT has collected and coded more than 31,000 categories, producing over 350 million quantitative variables and emotions that feed the GKG database, with updates every 15 minutes.

[4] About GDELT: The GDELT Project. Retrieved July 31, 2020.

Information extraction and preparation

The data-gathering process starts by extracting the information from digital sources such as newspapers, blogs, podcasts, analyst opinions, videos, and others, in more than 100 different languages.

Sources are cumulatively included in the system both manually by GDELT's team and automatically through GDELT's own crawler and Google News Feed. This continuous effort to expand the list of sources contributes to GDELT's increasingly comprehensive coverage of world events, but over time some sources may become obsolete since listed domains may change purpose or be reassigned, yet no source is delisted. Some domains in the list with seemingly unexpected names have been found to often be 'undercover' domains for news sites operating in countries where governments actively censor the web.[5]

Location of the source is extracted in three different ways: [6]

- From the TLD country information, such as ".uk" in a domain, when used;
- From domain registration information, which includes information on ownership and location;
- By creating a country-level histogram for each outlet that shows about which countries does each source report the majority of its time.

Moreover, as previously mentioned, all non-English news articles are machine translated into English before being processed and coded, which provides additional information. GDELT Translingual, the platform translating the news for GDELT, is said to be the largest real-time streaming news machine translation deployment in the world, translating news from over 65 languages.

This requires a significant amount of brute force in terms of processing power, and nowadays such a task only seems feasible using parallel computing ecosystems that can work with "MapReduce" protocols[7].

[5] All 'locationable' sources used by GKG can be found at http://data.gdeltproject.org/blog/2018-news-outlets-by-country-may2018-update/MASTER-GDELTDOMAINSBYCOUNTRY-MAY2018.TXT 'Non-locationable' sources are not included in this list, but they amount to a small percentage.

[6] Mapping the Media: A Geographic Lookup of GDELT's Sources: The GDELT Project. Retrieved July 31, 2020 from https://blog.gdeltproject.org/mapping-the-media-a-geographic-lookup-of-gdelts-sources/ .

[7] MapReduce is a programming system for distributed processing large-scale data in an efficient manner on a cloud.

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

Latent theme identification

To identify the theme behind each news piece, GDELT resorts to Natural Language Programming (NLP), which is necessary to infer topics from unstructured information. The key challenge at this stage is to address semantic problems related to "context", and this requires dealing with obstacles inherent to linguistics, i.e. polysemy, homonymy, synonymy, antonymy, hyponymy, and hypernym. [8] Such an inference process requires:

- Disambiguation, or linking of the meaning to a set of "interrelated conceptual meanings" (that is, a *context*), with the help of a lexicon such as "WORDNET LEXICON", with more than 120.000 groupings of related concepts ("*contexts*").
- An NLP technique to extract latent themes from unstructured information. The choice of NLP can be either a statistical technique such as the famous Dirichlet Localization Algorithms (ALD) or any other more complex -and more powerful-technique based on syntax, which identifies not only the equivalence of meaning of the tokenized words but also its position in a given syntactic construction (determinant of the meaning across all different languages).

Once the latent themes have been inferred from the unstructured data, they need to be supplemented with the necessary meta-information. This information may be related to the location, time, agent, or "URL" of the source, all of them necessary elements for the subsequent exploration of the extracted data.

Sentiment codification

Having identified the latent themes of a news piece and complemented them with its corresponding meta-information, GDELT proceeds to determine the underlying sentiment behind each article, based on its content. GDELT’s sentiment analysis uses Scherer's Type of Affective States (“STAS”), which reflects moods, emotions, interpersonal positions and attitudes towards a certain topic.

Mapping two different words linked to emotions from two different sets of synonyms is done with the help of a guide or sentiment dictionary. Dictionaries can be generic or have a specific purpose; some are general and others only refer to very specific areas (such as finance or crisis situations). To this date, the most famous sentiment

[8] Respectively: Same word, same root (etymology), different meanings; Same word, different root (etymology), different meanings; Different word, same meaning; Different word, opposite meaning; A phrase (or word) whose semantic field is more specific than its superordinate phrase (or word); A phrase (or word) whose semantic field is broader than its subordinate phrase (or word).

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

dictionary is Harvard University's "General Inquire IV".[9] The algorithm integrated in GDELT uses different sentiment dictionaries to assess the emotional charge of each concept or event analyzed in the database. GKG in particular uses its Global Content Analysis Measure (GCAM) algorithm to code sentiment, bringing together a whole array of leading content analysis tools (including the ones mentioned above) to capture over 2,230 latent dimensions, reporting density and value scores for each one of them.[10]

Tone coding

After identifying the positive and negative words in the document, GDELT calculates the overall tone of the text by calculating the positive and negative sentiment in it. The tone is a score resulting from the balance between positive and negative words divided by the sum of words in the text. These scores range from -100 (extremely negative) to +100 (extremely positive), but most frequently values range from -10 to +10, with 0 indicating neutral.

When the tone of a document is close to zero, it may point to one of two possibilities. Either (i) the text has a low emotional response or (ii) the positive and negative sentiment scores are approximately the same and they mutually cancel each other. Any of these situations can be detected by looking directly at the positive or negative scoring variable or the polarity variable.

Theme, actor and locations coding

Lastly, a labeling system categorizes the content of each topic extracted and analyzed. Labels are themes or categories that we find in different taxonomy glossaries such as the CrisisLex, World Bank Group[11] taxonomies or GDELT's own labeling system.[12]

If a theme begins with “tax_XXXX”, then “XXXX” is the literal word looked for. If it doesn’t, as in example “WB_XXXX”, then “XXXX” includes a list of keywords that are all coded with that same “XXXX” theme.

[9] There are others such as the "Linguistic Inquire Word Count", the "SentiWordNet", and the more event-based lexicon, known as the Goldstein Scale under the CAMEO taxonomy, for example.

[10] Introducing the Global Content Analysis Measures (GCAM): The GDELT Project. Retrieved July 31, 2020 from https://blog.gdeltproject.org/introducing-the-global-content-analysis-measures-gcam/ .

[11] https://vocabulary.worldbank.org/taxonomy.html

[12] The latest theme library can be found here (ranked, per the number following the theme, in terms of occurrences): http://data.gdeltproject.org/documentation/GKG-MASTER-THEMELIST.TXT

Funded by the Horizon 2020 Framework Programme of the European Union. Grant agreement number 822654

Similarly, actor coding occurs using actor and agent dictionaries.[13] The reliance on these dictionaries, however, which focus heavily on international politics, makes actor coding one of the most limited aspects of GDELT, since private sector, intergovernmental or non-government type actors simply are not included in the entity or actor dictionaries used.

Finally, every text is coded for location names in a similar way. The two dictionaries in use are the GEOnet Names Server (GNS), which includes US locations, and the Geographic Names Information System (GNIS), which covers the whole globe. [14]

Variables in the GKG database

After document processing, disambiguation, context identification -via the *latent theme identification*-, sentiment and tone detection and coding, each case -news piece- becomes a record in the database. The following information exists per record:

- GKGRECORDID
- DATE
- SourceCommonName
- DocumentIdentifier
- V2Themes: All themes found in the text.[15]
- V2Locations: All locations found in the text.[16]
- Entity extraction variables
    - V2Persons
    - V2Organizations
- V2Tone: Tone of text (according to GDELTs own sentiment dictionaries).[17]
- GCAM: All emotions found, and their relative strength, as found in the text according to a vast array of well-known external-to-GDELT sentiment dictionaries.[18]
- Dates: All dates found in the text.
- AllNames: all capitalized words (excluding first word following punctuation marks).

[13] See Chapter 3 *Actor Codebook* in Schrodt (2009), CAMEO Codebook.

[14] Fulltext Geocoding versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia: Leetaru, K. (2012). Retrieved July 31, 2020, from http://www.dlib.org/dlib/september12/leetaru/09leetaru.html .

[15] See above.

[16] See above.

[17] See *Tone coding* above.

[18] See *Sentiment coding* above.

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

METHODS

The potential use of GDELT in global governance seems very high. Not surprisingly, since its early presentation in 2011, various authors have explored the usability of GDELT for research purposes in the social sciences. It is worth exploring how GDELT has been used in academic research, which disciplines have used it, specifically which GDELT databases have been used, and how scholars have operationalized the data.

Article identification

We started with a very broad strategy of article identification. Our starting point was to explore all academic articles in the EBSCO database that contained the acronym "GDELT" anywhere in the text. This produced 59 academic papers in October 2019.

We read through the 59 articles with a special focus on the precise role GDELT played in the development of the author's thesis and on the overall methodology. Our aim at this stage was to distinguish which papers actually used GDELT data as opposed to those simply referencing it. After a second more thorough filtering in which the articles merely citing GDELT were discarded, we were left with 24 articles which, while heterogeneous in their fields of study and application of GDELT, had all integrated the database in the core of their methodology. From these 24, we left aside those that tackled purely methodological questions, resulting in a total of 19 articles.

This final sample comprised 19 papers published in English-language journals within a time frame ranging from 2014 to 2019.

As noted earlier, after the first filtering, 5 articles were discarded due to their methodological nature. That is, because far from using GDELT to operationalize the IV, the DV or both, the articles discussed and examined the database per se, exploring its functioning and accuracy, highlighting its weaknesses and suggesting future improvements.

Exploring the functioning and accuracy of GDELT, Yuan, Liu and Wei (2017:03), *Exploring inter-country connection in mass media: A case study of China*, in which the authors, citing Ward et al. (2013), state:

> "Another study by [Ward et al. (2013)] uses a commercial dataset - the Worldwide Integrated Crisis Early Warning System (ICEWS) - to validate the GDELT dataset on temporal trajectories at a fixed location. The authors found that GDELT and ICEWS both captured major events in Egypt between 2011 and 2012, but GDELT appears to cover more events than ICEWS".

Yuan, Liu, and Wei also note that:

> "The study by Arva et al. (2013) cross-compared ICEWS and GDELT at the country level, and they reported that the GDELT data performs as well or better

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

> than the data in the original ICEWS dataset in detecting hotspots of conflict at the country level."

These authors also conclude: "[Our analysis confirmed] the geocoding accuracy of GDELT at the national level, as well as the reliability of GDELT in modeling connection between countries by looking into countries related to China" (Yuan, Liu, Wei; 2017:04)

In another article by Dos Santos et al. (2016), published in the Journal GeoInformatica under the title *The big data of violent events: algorithms for association analysis using spatio-temporal storytelling*, the authors aim at evaluating different storytelling methods and their effectiveness in uncovering real-world storylines for exploratory analysis. After conducting evaluations on the basis of three methods (Distance-based Bayesian Interference (DbB), Spatial Association Index (SAI), and Spatio-logical Inference (SLI)), the study concludes that spatio-temporal storytelling is able to capture important associations among violent events reported in social media and news-based databases (such as GDELT), two common sources of Big Data (Dos Santos et al., 2016:28).

Another discarded paper proposes a possible future improvement of GDELT related to translation. The authors use Turkish language news to show how GDELT's translation has room for improvement: "We first constructed a large database of synonym and antonym word pairs from the existing lexicon and publicly available Turkish language sources. We then modeled this database as a graph, and finally, we propagated tone and polarity values from existing words to newly added words. Upon testing this extended 49K word lexicon using manually labeled test corpus, we obtained significant accuracy increases." (Sağlam, Genç & Sever, 2019:02)

ARTICLE ANALYSIS

Benefits, challenges & reliability of GDELT

The analyzed papers used GDELT for several reasons. Firstly, it is a live ongoing project that continuously collects news, coding all-new news articles into the system with updates every 15 minutes. Moreover, the GDELT Event database goes back four decades, while GKG goes back just to 2014. Levin, Ali, and Crandall's (2018) study of conflict events during the Arab spring finds that "GDELT events reflected those protests, assaults, and fights immediately, without lag time, thus providing high correspondence with actual conflicts and death events."

Secondly, GDELT's data covers the entire globe, since it gathers and codes news sources from all countries. For countries subject to government censorship, such as China, it only covers media outlets which are open to the world, such as the Communist

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

Party's mouthpieces, Global Times and others. These two features are well portrayed in Qiao (2017). They use the GDELT to gather news from Cambodia, Malaysia, Indonesia, Philippines, and Thailand from 2001 to 2016. Similarly, Levin (2018) uses GDELT to analyze news from Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria and Tunisia, from 2000 to 2016.

In any case, GDELT has also been used for single-country studies, as it has been the case for Spain, Nigeria, Syria, Egypt, or South Africa.

The reliability of GDELT

Nine papers mention and discuss reliability issues of GDELT in their respective methodology sections.

Bunte (2016) finds that GDELT doesn't capture some interreligious violent events in Nigeria because it relies too heavily on digital newspapers' reporting, and not all events make it to those outlets.

Particularly regarding the GDELT Event database, identifying an event and aggregating different news records reporting on the same event is among the greatest challenges GDELT encounters. The key problem in this respect is to distinguish whether two news pieces refer to the same event or two separate events. While in Vargo (2016) two coders hand-reviewed 286 GDELT coded events and found an 80% reliability, Dos Santos review of three quarter million records finds that GDELT identified about 1000 events out of over 2500 known events, with a precision rate around 40%. Jenkins & Maher (2016) finds that after keyword and temporal filtering, de-duplication of events, and machine-learning classification of real events from non-events or planned events, only 21% of GDELT's valid URLs indicate a true protest event. As Jenkins and Maher (2016), state: "A naive user [of GDELT] might take literally the 649 kidnappings reported for Nigeria during the month after April 14, 2014. Actually, this is the number of news reports about the same mass kidnapping by the Boko Haram that GDELT located" (Jenkins & Maher, 2016:50).

Furthermore, in order to maximize its reliability, GDELT has sought to expand its boundaries in terms of news-gathering. Over time, GDELT has worked to diversify its events sources to increasingly include non-Western sources on its list. As one of the methodological papers reviewed--*Growing pains for global monitoring of societal events in the Journal Science* by Wang et al. (2016:02)--states: "[Despite] GDELT has expanded the number and variety of its sources...—incorporating, for example, more non-Western news sources[--]reliance on English-language news coverage results in stronger correlations for states more often in the Western press (e.g., Brazil)" (Wang et al., 2016:03).

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

Disciplines using GDELT

From our sample of 19 academic articles, Information and Communication Science is the discipline with the most papers published using GDELT: 6 in journals such as *Information Services & Use, New Media & Society, Journal of Communication*. Coming in a close second, Conflict Studies has 5 publications (published in *Journal of Peace Research, Journal of Conflict Resolution, Civil Wars*, and others).

Papers focusing on methodological issues of big data —such as Wang et al's (2016) cited above--added up to 5 in total (published in *Intelligent Data Analysis, Science, GeoInformatica, Computers, Environment & Urban Systems*). Political Science journals (*International Interactions, International Political Science Review & Journal of Development Studies*) published 3 articles, and Finance and Business (*International Journal of Engineering Business Management & Review of Financial Studies*) and Sociology (*Current Sociology, Discrete Dynamics in Nature & Society*) each published two articles. Lastly, only one paper has been published in the field of Geography.

Discipline	Journal	No. of articles
Information	Information Services & Use*	1
	International Journal of Interactive Multimedia & Artificial Intelligence	1
	Journal of Communication	1
	Journal of Information Science	1
	Journalism & Mass Communication Quarterly	1
	New Media & Society	1
Conflict	Civil Wars	1
	Conflict Management & Peace Science	1
	Journal of Conflict Resolution	1
	Journal of Peace Research	2
Political Science	International Interactions	1
	International Political Science Review	1
	Journal of Development Studies	1
Finance, Economics & Business	International Journal of Engineering Business Management	1
	Review of Financial Studies	1
Sociology	Current Sociology	1
	Discrete Dynamics in Nature & Society	1

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

Geography	Applied Geography	1
Methodology	Computers, Environment & Urban Systems	1
	Science	1
	GeoInformatica	2
	Intelligent Data Analysis	1

Figure 2. Journals and disciplines represented in our final sample.

Geography covered

From our 19 articles, the geographic coverage per field is as follows:

- **Information and Communication:** In this field, the main geographic focus is the USA, being present in three out of the six papers. Two out of the three papers left set their focus of study at a world-level, and the remaining one sets its scope on Spain.
- **Conflict:** Geographic coverage of conflict studies papers are China, India, MENA, Nigeria, Syria, Philippines and Thailand.
- **Methods:** The countries and regions covered by methodological papers using GDELT are China, Mexico, Middle East, Asia, USA and Latin America.
- **Political Science:** The articles using GDELT in this field cover South Africa, Europe and the Amazon and Jordan River basins.
- **Sociology:** The papers studied within the sociology spectrum have their focus in South East Asia, and to a minor extent, in Europe.
- **Business and Finance:** Egypt and the top 10 EU countries (in terms of GDP).
- **Geography:** In this field, the sole paper published using GDELT studies the MENA region.

Overall, conflict studies seem to use GDELT for non-Western countries and regions, while information studies focus on the Western world. The other disciplines appear to be geographically more diverse.

GDELT data usage

Among the 19 articles studied, 7 use the database to operationalize the dependent variable (DV), another 7 times the independent variable (IV), and an additional 4 studies both the dependent and independent variable.[19] For example, Bedasso (2016) uses GDELT to construct its DV of protests and explore the effects of income. Bunte (2016) does so to construct a measure of interreligious violence (their DV) and explore whether power-sharing agreements decrease the former. Elshendy (2017) constructs

[19] Leetaru (2015) *Mining libraries: Lessons learned from 20 years of massive computing on the world's information* was found to not utilize GDELT data either as dependent or independent variables.

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

one of its independent variables (tone of economic news) using GDELT and explores how it affects confidence indexes. Other studies use GDELT data for variables on both sides of the equation. These have to do with predictive model building (Dos Santos, 2016) or with social network analysis (Vargo, 2018).

GDELT may be used to research different units of analysis. It can analyze relationships (between themes and entities), a country, a given policy, issue or sector, and/or an event. For instance, Bedasso (2016) uses GDELT to measure the annual level and intensity of protests in South Africa between 1979 and 2012. Yuan (2017) uses GDELT to measure political relations between China and other countries from 2011 and 2017. Other scholars measure protests and events at district level (Bunte, 2016) or subdistricts (De Juan, 2015). Yet other authors use GDELT to measure volume and tone of news regarding a theme in one country (Elshendy, 2016), or the centrality of a given theme in the network of relationships between themes in the news published by major media groups (where relations are co-occurrences in the same news piece) (Vargo, 2018).

in relation to the GDELT Project's databases used, i.e. GDELT Events and GDELT GKG, most papers use the GDELT Events database. These papers mostly create a time-series variable from aggregating counts of an event (or type of events, i.e. protests, strikes, military assaults…) in a given space/territory.

In terms of analysis, some explore the data qualitatively, some others use correlations, but the majority of articles use regression analysis.

Discussion

What makes GDELT very attractive is its extensive coverage of world events, collected and coded every 15 minutes. The GDELT Event database goes back four decades, while GKG goes back to 2014. Furthermore, the GDELT's world coverage allows for its application to analyze phenomena in different regions of the world. Despite the prevalence of the disciplines of Information and Conflict, GDELT data is utilized in a vast array of research fields, both as direct and indirect variables - sometimes even to construct both of them. Methodologically, time-series variables for aggregated event counts seem to be among the preferred methods of operationalization, later to be analyzed qualitatively or quantitatively.

Funded by the Horizon 2020 Framework Programme of the European Union. Grant agreement number 822654

A DETAILED LOOK AT THE SPECIFIC USE OF GDELT-BASED MEASURES

- Bedasso qualitatively analyzes the temporal evolution of protests in South Africa. They use GDELT Events to calculate the level and intensity of protest between 1979 and 2012 (GDELT Project, 2014). They take the count of events coded under the 'protest' category for every year and divided it by the total number of events in that particular year to get a "protest ratio".
- De Juan uses GDELT Events to draw up a time series of annual counts of violent events in Thailand and the Philippines. The authors qualitatively analyze the two time series.
- De Juan explores whether governments with less public services have greater levels of violence. In one of their logit models (used as robustness check), they use a count of GDELT Events as dependent variable, being their main IV the population's access to electricity.
- Levin creates a time series of news items related to conflicts based on GDELT Events data and compares it to actual conflicts recorded. All variables were aggregated and analyzed as monthly time series at country level (n=18), covering all Arab countries (including the Palestinian Authority) from Morocco to the entire Arab Peninsula. They carry out Spearman's rank correlations for each variable to examine its changes during different time periods, as well as over the entire time frame.
- Quiao uses GDELT Events to build and train a Hidden Markov Models (HMMs) based framework to predict the number of events in a given country based on the number of previous events, type of previous events, and number of mentions of previous events.

A slight variant of the above is to use GDELT Events data as variable but selecting not where the events (or types of events) occurred but which two -or more- entities did they involve.

- Bunte builds his dependent variable for interreligious communal violence using the GDELT Events database. To capture the interreligious dimension, they only count events where the adversaries were clearly identified as Christians and Muslims. They then run a negative binomial count model against the violence count variable. The independent variables include welfare quintile level of the community, the district-level power-sharing binary variable (i.e. inclusive or not), a district-level propensity score of a violent event being reported or not by news media, and some other control variables.
- In Morlino's study, unconventional political participation (as a measure of quality of democracy) is measured by using a count of protest actions –demonstrations, strikes, boycotts and riots– initiated by civilians, the opposition, labor organizations, the media, non-governmental organizations and human rights

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

organizations. A 25 country*year cross-sectional time-series is created and linearly regressed as a DV against several other economic time series.

- Dale uses GDELT conflict/cooperation events in the Amazon River Basin to compare how the different basin countries respond to each other. Their paper is about identifying the right aggregation scale (daily, weekly…) for events to then be able to detect stimulus-response relationships between the actors involved in the events and showing how aggregation scales have important effects on a study.

A few other papers use GDELT Events to construct more elaborate variables:

- Elshendy constructs a network from each dyad (and corresponding location) identified in GDELT events. They focus on 10 EU countries and business events. Based on the network, they produce independent variables such as different centrality and centrality oscillations scores. They then run a multilevel regression on country confidence indices (CCI, BCI as dependent variables).
- Davis measures the political relations between China and India and third countries using GDELT Events. They sum the severity-weighted number of negative events to create a single annual observation and take the log to smooth the distribution for each bilateral relation. They restrict the negative event counts to actions performed by a partner country upon China and India, respectively. They then do a regression on the imports where one of the IV is a one-year lagged political relation measure between the importing country (China or India) and the exporting country. Their analysis covers the timespan 1993-2012.
- Acemoglu uses GDELT Events data set to construct a measure of the sensitivity of a firm’s stock return to general unrest in the country. They use this data set to obtain a list of strikes, boycotts, riots, and instances of ethnic clashes between Muslims and Christians that occurred between January 1, 2005, and December 31, 2010. They then regress the stock returns for each firm (i) on a dummy variable that is one on the two trading days following one of the events on our list, and refer to the slope coefficient of this regression as “unrest beta, β”. They then use the sensitivity measure as a control variable in an OLS regression between cumulative return of firms and a firm’s connections (military, political party, Islamic).

Three papers use the GDELT GKG database:

- Using GDELT GKG, Elshendy coded two variables: The daily number of newspaper articles covering the ‘WTI Crude Oil Price’ search criterion and the count of all oil-industry-related organization names or advisory councils mentioned all over the world during the period of the study. A Pearson correlation analysis was used to test the association between the social

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

awareness variables (including the two GDELT based ones) and the WTI crude oil price on the first, second and third lags. They also used Granger causality statistical tests to identify if one time series provided useful information which could help in forecasting another time series (Table 3).

- Vargo et al. (2018) and Guo and Vargo (2017) (in two articles focusing on different media types) use GDELT GKG to create times series of issues (16) per media type (9). The time series of centrality scores (issue centrality score × media type × day) were constructed by linking issues if they were covered in the same article. Granger causality models were constructed for each issue*media type time series. Running F-tests provided values of significance in which Granger causality could be determined.
- Vargo et al. create a global salience index (GSI) of a country defined as the percentage of news items that mentioned it (not including items from its own news sites) out of all news items that mentioned any country name (e.g., the percentage of non-American news items mentioning the United States out of all non-American news items that mentioned countries). Correlation tests were conducted to explore the relationship between a country's position in the world system (1-core; 2-semiperipheral; 3-peripheral)[20] and its GSI score (H1); and a country's GDP (worldbank.org), expressed in dollars, and its GSI score (H2). Additionally, to determine how the media in different countries interacted with each other they employed Granger causality tests: they do an ordinary least squares (OLS) regression, a one-day time lag was tested through the regression of each country/issue media agenda.[21]

CONCLUSION

In this paper we have analysed how GDELT has been used in academic research. GDELT's objective is to register and code as many world events as possible and the sentiment associated with their coverage. This article has provided an ample review on the usage that scholars have given to the database and its multiple perks in global governance and crisis prediction. In order to carry out our review, we started with the analysis of 59 academic articles that mentioned GDELT at some point. After discarding those papers that were only citing the database as an example, and later proceeding

[20] The study selected an initial list of core, peripheral, and semiperipheral countries based on Chase-Dunn, Kawana, and Brewer (2000) and Babones (2005).

[21] Series X is said to "Granger cause" another time series Y if regressing for Y in terms of past values of both X and Y results in a bettermodel for Y than regressing only on past values of Y

Funded by the Horizon 2020 Framework Programme of the European Union.
Grant agreement number 822654

to discard those that were purely methodological,[22] we were left with a total of 19 articles for the posterior analysis.

First, we laid down the foundations of GDELT. The database collects and codes more than 31,000 categories, producing over 350 million quantitative variables and emotions, with updates every 15 minutes. Furthermore, through the study of GDELT Events and GDELT GKG, we have seen how the gathered data is later classified, allowing for a precise and narrow categorization of the information, crucial for precise sentiment analysis. The data is coded according to theme, sentiment, tone, actors and locations.

Second, the core part of the article has reviewed the usage that scholars have given to GDELT Events and GDELT GKG throughout the 19 aforementioned papers. We have structured our examination into multiple categories. The first one, the benefits and challenges that GDELT presents, alongside its reliability. From the surveyed articles we have inferred that the main advantage that both databases present is its vast array in terms of news coverage. This perk is twofold, first on the grounds of timespan, where GDELT goes back 40 years into news analysis. Second, on the grounds of coverage and translation, gathering news from over 100 countries in more than 60 languages. These two factors, along with its constant content update and its gradual inclusion of non-Western newspapers, have earned GDELT a positive degree of reliability among scholars. The second category that we have reviewed is the disciplines covered. Data shows that the main field where GDELT Events and GDELT GKG have been used is Information and Communication Science, with a total of 6 publications out of the 19 papers studied, closely followed by Conflict Studies, with a total of 5 publications. Furthermore, the papers analysed have a broad geographic scope, covering events at a worldwide level. Overall, Conflict Studies use GDELT for non-Western countries and regions, especially for South-East Asia and the MENA, while Information Studies focus primarily on the Western world. The other disciplines appear to be geographically more diverse. Regarding the particular usage that scholars have given to GDELT Events and GDELT GKG, it is used to operationalize the dependent and independent variable. Among the 19 articles studied, 7 have used the database to operationalize the dependent variable (DV), 7 have used it for operationalizing the independent variable (IV), and an additional 4 studies have used it both for the dependent and independent variable (DV/IV). This further proves that the database is a versatile tool, giving a plethora of opportunities.

Third, we can briefly summarize the possible future improvements to the databases covered in this text. The first issue encountered makes reference to the amelioration in translations once GDELT collects news from non-English sources. The second issue

[22] These papers were discarded because they analyzed the reliability of GDELT and its capacity to predict the emergence of violent events worldwide, yet the database was not used as a tool to test a particular hypothesis.

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

comes in terms of bias, as there is a tendency from the database to collect news from Western sources. Nevertheless, this issue has been gradually corrected, as recently the database has worked to include news sources from the Global South. The third issue, as mentioned in *The reliability of GDELT* section, namely identifying an event and aggregating different news records reporting on the same event, is among the greatest challenges GDELT Events encounters. Lastly, we find it relevant that the database only includes news from Chinese newspapers that have world wide web exposition, and thus, lacks access to those that are only available on the Chinese intranet. This, in face of future extrapolation, may present a significant defiance for the database and its reliability given the importance of China in world events.

In sum, GDELT, and particularly GDELT Events and GDELT GKG have proved to be a resourceful and multifaceted tool with highly precise prediction indexes. While the usage that scholars have given to it is still limited, there is great potential to exploit its many applications, from information and communication to conflict study and global governance. As it has been stated throughout the paper, the potential of GDELT to contribute to research through its sentiment analysis is remarkable. Nonetheless, we have found that it is still in an embryo stage, and that scholars have not yet developed and ventured its full potential.

ANNEX I. ARTICLES REVIEW

Mining libraries: Lessons learned from 20 years of massive computing on the world's information

Leetaru, Kalev 2015

Information Services & Use / *Information*

Abstract: Over the course of the last two decades I have explored the informational undercurrents of the world's information and the potential of mass data mining of libraries through a myriad of lenses, both technical and methodological. From founding my first Internet startup twenty years ago as an eighth grade middle school student, to running one of the world's largest global monitoring platforms today, my work has debuted a myriad of new datasets, methodologies, and scales to the study of how we understand our global world. A central theme of that work has been around how creative "reimagining" of information through the emerging world of massive computing power can offer powerful and unexpected new lenses onto the world around us, and the incredible future that awaits as libraries transition from being museums of artifacts to becoming conveners of information and innovation that empower a new era of access and understanding of our world. This paper offers a unique detailed view of what it looks like deep in the trenches of the data revolution, surveying a selection of my projects over the last two decades and the lessons they offer libraries and publishers moving forward into our data-driven future.

Main finding: Brief review of the GDELT's purpose and functioning (with a particular incise on what does it analyze), and description of the GDELT 2.0 launched in 2015

Database	Events
IV/DV	N/A
Variables	N/A
Measure	N/A
Locations	N/A
Analysis	Descriptive/Regression
Period	1995-2015

Funded by the Horizon 2020 Framework Programme of the European Union. Grant agreement number 822654

Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy

Bodas-Sagi, Diego J.

2015

International Journal of Interactive Multimedia & Artificial Intelligence / *Information*

Abstract: The growing demand for affordable, reliable, domestically sourced, and low-carbon electricity is a matter of concern and it is driven by several causes including public policy priorities. Policy objectives and new technologies are changing wholesale market design. The analysis of different aspects of energy markets is increasingly on the agendas of academics, firms' managers or policy makers. Some concerns are global and are related to the evolution of climate change phenomena. Others are regional or national and they strongly appear in countries like Spain with a high dependence on foreign energy sources and high potential of domestic renewable energy sources. We can find a relevant case in Spanish solar energy policy. A series of regulatory reforms since 2010 reduce revenues to existing renewable power generators and they end up the previous system of support to new renewable generation. This policy change has altered the composition of the energy market affecting investment decisions. In this paper, we analyze the public opinion about energy policy of the Spanish Government using the Global Database of Events, Language, and Tone (GDELT). The GDELT Project consists of over a quarter-billion event records in over 300 categories covering the entire world from 1979 to present, along with a massive network diagram connecting every person, organization, location, and theme to this event database. Our aim is to build sentiment indicators arising from this source of information and, in a final step, evaluate if positive and negative indexes have any effect on the evolution of key market variables as prices and demand.

Main finding: We detect negative feelings about the solar energy policy introduced by the Spanish government in 2015. On the other, hand, we find weak correlation between the indexes (tone) in mentions from GKG database and average daily log prices of energy. We do not find any correlation to average daily energy demand.

Database	Events
IV/DV	IV
Variables	Public opinion re: energy policy
Measure	Country-level
Locations	Spain
Analysis	Correlation
Period	2015

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

Global Intermedia Agenda Setting: A Big Data Analysis of International News Flow

Guo, Lei

2017

Journal of Communication / *Information*

Abstract: This study contributes to international news flow literature methodologically, by significantly expanding its scope, and theoretically, by incorporating intermedia agenda-setting theory, through which we reveal how news media in different countries influence each other in covering international news. With a big data analysis of 4,708 online news sources from 67 countries in 2015, the study shows that wealthier countries not only continue to attract most of the world news attention, they are also more likely to decide how other countries perceive the world. However, international news flow is not as hierarchical and U.S.-centric as found earlier. Online-only, emerging media in core countries are not necessarily more impactful in setting the world news agenda than those in (semi)peripheral countries.

Main finding: Wealthier countries not only continue to attract most of the world news attention, they are also more likely to decide how other countries perceive the world. However, international news flow is not as hierarchical and U.S.-centric as found earlier.

Database	GKG
IV/DV	N/A
Variables	Relationship between themes
Measure	Countries
Locations	Worldwide: 67 countries
Analysis	Correlations with economic variables & F tests of Granger causality tests
Period	2015

Using four different online media sources to forecast the crude oil price

Journal of Information Science / *Information*

Abstract: This study looks for signals of economic awareness on online social media and tests their significance in economic predictions. The study analyses, over a period of 2 years, the relationship between the West Texas Intermediate daily crude oil price and multiple predictors extracted from Twitter; Google Trends; Wikipedia; and the Global Data on Events, Location and Tone (GDELT) database. Semantic analysis is applied to study the sentiment, emotionality and complexity of the language used. Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) models are used to make predictions and to confirm the value of the study variables. Results show that the combined analysis of the four media platforms carries valuable information in making financial forecasting. Twitter language complexity, GDELT number of articles and Wikipedia page reads have the highest predictive power. This study also allows a comparison of the different fore-sighting abilities of each platform, in terms of how many days ahead a platform can predict a price movement before it happens. In comparison with previous work, more media sources and more dimensions of the interaction and of the language used are combined in a joint analysis.

Main finding: Four media platforms can deliver valuable information at different 'speeds': Twitter is most informative at a 1-day lag, GDELT and Wikipedia at a 2-day lag, and Google Trends at a 3-day lag.

Elshendy, Mohammed	2018
Database	GKG
IV/DV	IV
Variables	News events
Measure	GDELT number of articles related to oil
Locations	USA
Analysis	Pearson correlation with social awareness variables
Period	2013-2015

Networks, Big Data, and Intermedia Agenda Setting: An Analysis of Traditional, Partisan, and Emerging Online U.S. News

Vargo, Chris J.

2017

Journalism & Mass Communication Quarterly / *Information*

Abstract: This large-scale intermedia agenda–setting analysis examines U.S. online media sources for 2015. The network agenda–setting model showed that media agendas were highly homogeneous and reciprocal. Online partisan media played a leading role in the entire media agenda. Two elite newspapers—The New York Times and The Washington Post—were found to no longer be in control of the news agenda and were more likely to follow online partisan media. This article provides evidence for a nuanced view of the network agenda–setting model; intermedia agenda–setting effects varied by media type, issue type, and time periods.

Main finding: Network agendas of various media outlets are highly interdependent, symbiotically networked and homogeneous.

Database	GKG
IV/DV	-
Variables	Relationship between themes
Measure	Global Salience Index (GSI) of a country is the percentage of news items that mentioned it (not including items from its own news sites) out of all news items that mentioned any country.
Locations	USA
Analysis	Correlation tests w economic measures & Granger causality tests: OLS regression
Period	2015

Funded by the Horizon 2020 Framework Programme of the European Union.
Grant agreement number 822654

The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 201.

New Media & Society / *Information*

Abstract: This study examines the agenda-setting power of fake news and fact-checkers who fight them through a computational look at the online mediascape from 2014 to 2016. Although our study confirms that content from fake news websites is increasing, these sites do not exert excessive power. Instead, fake news has an intricately entwined relationship with online partisan media, both responding and setting its issue agenda. In 2016, partisan media appeared to be especially susceptible to the agendas of fake news, perhaps due to the election. Emerging news media are also responsive to the agendas of fake news, but to a lesser degree. Fake news coverage itself is diverging and becoming more autonomous topically. While fact-checkers are autonomous in their selection of issues to cover, they were not influential in determining the agenda of news media overall, and their influence appears to be declining, illustrating the difficulties fact-checkers face in disseminating their corrections.

Main finding: Fake news websites are increasing, but do not exert excessive power. Instead, fake news has an intricately entwined relationship with online partisan media, both responding and setting its issue agenda. Fact-checkers were not influential in determining the agenda of news media.

Vargo, Chris J	2018
Database	GKG
IV/DV	-
Variables	Centrality of issue (i.e. Themes) per media type
Measure	Time series (issue centrality score × media type × day), where centrality was derived from an issue co-ocurrance network
Locations	USA
Analysis	Correlations with economic variables & F tests of Granger causality tests
Period	2014-2016

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

Framing Political Violence: Success and Failure of Religious Mobilization in the Philippines and Thailand

De Juan, Alexander

2015

Civil Wars / *Conflict*

Abstract: How do religious civil wars evolve? Many violent conflicts are fought between groups of different faiths. The paper argues, however, that religious differences rarely directly lead to conflict onset. Rather, the apparent religious dimension of many civil wars is a consequence of successful religious framing. Political and military leaders offer religious interpretations designed to legitimize the use of force and to mobilize believers to violent action. Such framing processes can be more or less successful, depending inter alia on the authority of the political and religious leadership, on the coherence and appropriateness of the frames, on the existence of persuasive counter-frames, and on the availability of communication infrastructures that allow for effective dissemination of religious frames. Comparing violent conflicts in the Philippines and Thailand, the paper shows that religious mobilization can fail along the theoretically predicted lines.

Main finding: Comparing violent conflicts in the Philippines and Thailand, the paper shows that religious mobilization can fail along the theoretically predicted lines.

Database	Events
IV/DV	DV
Variables	Violent events
Measure	A time series of annual counts of violent events at province level in Thailand and Philippines
Locations	Philippines & Thailand
Analysis	Qualitative exploration
Period	1979-1988

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

Identity-based political inequality and protest: The dynamic relationship between political power and protest in the Middle East and North Africa

Bodnaruk Jazayeri, Karen

2016

Conflict Management & Peace Science / *Conflict*

Abstract: In this study, I evaluate the effect of identity-based political inequalities on the probability of nonviolent and violent resistance in the Middle East and North Africa (MENA). During the Arab Spring, many states with higher levels of ethnic and religious political inequalities experienced lengthy resistance movements and violence, while states with lower levels of political inequalities largely experienced less volatility. I find support for my expectations using data spanning 1960–2011. Not only do results suggest that in MENA states with higher levels of identity-based political inequalities experience more conflict, but results from a global analysis suggest that this relationship is particularly robust within MENA. Sub-Saharan Africa is also conflict prone; however, the onset of nonviolent and violent movements is lower than MENA.

Database	Events
IV/DV	DV
Variables	Social unrest
Measure	N/A
Locations	MENA
Analysis	N/A
Period	N/A

Main finding: Because composite events are episodically aggregated, the temporal aggregation issues are different from those of “atomic” event data sets.

State Control and the Effects of Foreign Relations on Bilateral Trade

Davis, Christina L. 2019

Journal of Conflict Resolution */ Conflict*

Abstract: Can governments still use trade to reward and punish partner countries? While World Trade Organization (WTO) rules and the pressures of globalization restrict states' capacity to manipulate trade policies, politicization of trade is likely to occur where governments intervene in markets. We examine state ownership of firms as one tool of government control. Taking China and India as examples, we use new data on bilateral trade disaggregated by firm ownership type as well as measures of political relations based on bilateral events and United Nations voting data to estimate the effect of political relations on import flows since the early 1990s. Our results support the hypothesis that imports controlled by state-owned enterprises are more responsive to political relations than imports controlled by private enterprises. This finding suggests that politicized import decisions will increase as countries with partially state-controlled economies gain strength in the global economy. Extending our analysis to exports for comparison, we find a similar pattern for Indian but not for Chinese exports and offer potential explanations for these differential findings.

Main finding: Where governments maintain control over the economy (i.e. SOEs), trade continues to follow the flag.

Database	Events
IV/DV	IV
Variables	Political relations between countries
Measure	Yearly quality of relationships between countries
Locations	China, India, RoW
Analysis	Regression: Gravity model of trade
Period	1990-2011

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

Local power-sharing institutions and interreligious violence in Nigeria

Bunte, Jonas B. 2016

Journal of Peace Research / *Conflict*

Abstract: News reports of clashes between Muslims and Christians in countries such as Nigeria are increasingly common. Yet, interreligious violence erupts only in some communities but not others. Under what conditions does religious identity become the fault line of communal violence? We argue that informal power-sharing institutions on the communal level are essential in shaping the incentives of potential perpetrators. We provide both qualitative and quantitative evidence for our claim that districts in which informal power-sharing agreements exist are less likely to experience interreligious violence. We conducted interviews with community leaders in 38 Nigerian districts to trace the process by which local power-sharing institutions exert influence on actors' incentives to engage in religious violence. We complement this with quantitative analyses of a new dataset capturing interreligious violence on a subnational level. The analyses show that the overall degree of interreligious violence is significantly lower in districts with power-sharing than in those without. We also identify two causal mechanisms through which informal power-sharing institutions operate. First, these institutions affect the incentives of elites to appeal for cooperation. We show that the rhetoric of elites in districts with power-sharing is significantly more conciliatory. Second, power-sharing affects the general population's perception of the interreligious tensions. Individuals living in districts with power-sharing institutions are less likely to experience religious diversity as threatening. Local-level informal power-sharing institutions are therefore an important foundation for communal peace and interreligious cooperation.

Main finding: Individuals living in districts with power-sharing institutions are less likely to experience religious diversity as threatening.

Database	Events
IV/DV	DV
Variables	Violent events
Measure	Count events at district level where the two adversaries were clearly identified as Christians and Muslims.
Locations	Nigeria
Analysis	Regression: Negative binomial count model
Period	2000-2006

Funded by the Horizon 2020 Framework Programme of the European Union. Grant agreement number 822654

The Ba‘athist blackout? Selective goods provision and political violence in the Syrian civil war

De Juan, Alexander

2015

Journal of Peace Research / *Conflict*

Abstract: Many authoritarian regimes selectively provide critical segments of the population with privileged access to goods and services, expecting political support in return. This article is interested in the effects of this regime strategy: Is violent opposition less likely to occur in subnational regions bound to the ruling elite through such patron–client networks? For its empirical analysis, the article makes use of crowdsourcing data on the number and geospatial distribution of fatalities in the Syrian civil war from March 2011 to November 2012. In terms of selective goods provision, the focus is on the electricity sector. Satellite images of the earth at night are used to proxy spatial variations in the public distribution of electricity in times of power shortages. These data are complemented with information from the last Syrian population census of 2004. Estimations from fixed effect logit models lend support to the hypothesis that the risk of violence has been lower in subdistricts that have been favored by the ruling regime in terms of preferential access to material goods. This hypothesis is further corroborated with qualitative evidence from Syrian localities.

Main finding: The risk of violence has been lower in subdistricts that have been favored by the ruling regime in terms of preferential access to material goods.

Database	Events
IV/DV	DV
Variables	Violent events
Measure	Count of GDELT events at subnational regions
Locations	Syria
Analysis	Fixed effects logit models
Period	2011-2012

Minimizing the Effects of Temporal Aggregation on Event Data Analysis

Dale, Thomas G.

2014

International Interactions / *Political Science*

Abstract: Event data remains one of the best means for evaluating reciprocity and triangularity in international politics. One difficulty with using this type of data has been its susceptibility to the statistical effects of temporal aggregation. This article examines the concept of the natural interval in event data analysis, specifies how a user-selected aggregation interval affects measured stimulus–response behavior, and proposes a method for calculating a minimum threshold for the natural interval. The article then examines how such a minimum threshold reduces the impact of misspecification on perceived relationships for the Amazon River Basin and the Jordan River Basin.

Main finding: Taking the system harmonic mean of the minimum expected natural interval for each dyad clearly improves overall model validity at the system level as well.

Database	Events
IV/DV	DV/IV
Variables	Political relations between countries
Measure	Cooperation/Conflict among countries sharing river basins
Locations	Amazon River Basin and the Jordan River Basin
Analysis	Regression
Period	N/A

What is the impact of the economic crisis on democracy? Evidence from Europe

Morlino, Leonardo

2016

International Political Science Review */ Political Science*

Abstract: There are already a number of good accounts of the impact of the recent 2008–2014 economic crisis on European democracies. However, no systematic assessments of how it has affected specific aspects of democracy have so far been carried out. We explore its impact on European democracies in several areas by adopting the ‘quality of democracy’ framework. Our analysis shows that the measures we employ capture the variation in quality during this ‘troubled’ period. The empirical analysis suggests that a shrinking of private and public resources due to an economic downturn prompts three reactions: (a) a general deterioration of the rule of law; (b) citizens become more sensitive about what governments deliver; and (c) detachment from the institutional channels of representation along with a choice to protest.

Database	Events
IV/DV	DV
Variables	Social unrest
Measure	Protest, year, country
Locations	10 EU countries
Analysis	Linear Regression
Period	2000-2014

Main finding: How [GFC] has affected specific aspects of democracy.

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

A Dream Deferred: The Microfoundations of Direct Political Action in Pre- and Post-democratisation South Africa

Bedasso, Obikili 2016

Journal of Development Studies / *Political Science*

Abstract: The stability of young democracies may be threatened by persistent protest as the economic legacies of the old autocratic regimes tend to outlive the defunct political structures. This paper seeks to explore the micro-level predictors of protest potential in South Africa before and after the end of apartheid. The results of the cohort analysis reveal that the political consciousness of the anti-apartheid struggle has a lasting effect. The gap between actual income and expected returns to education explains protest potential better than comparison of one's income with that of a reference group. The effect of race on protest potential has diminished over time.

Main finding: Protests and political action are on the rise, highly linked with the economic situation. There is a decrease just after the end of the Apartheid system.

Database	Events
IV/DV	DV
Variables	Social protests
Measure	Count of events coded under the ‘protest’ category for every year and divided it by the total number of events in that particular year
Locations	South Africa
Analysis	Qualitative exploration
Period	1979-2012

Funded by the Horizon 2020 Framework
Programme of the European Union.
Grant agreement number 822654

Big data analysis of economic news

International Journal of Engineering Business Management / *Finance*

Abstract: We propose a novel method to improve the forecast of macroeconomic indicators based on social network and semantic analysis techniques. In particular, we explore variables extracted from the Global Database of Events, Language, and Tone, which monitors the world's broadcast, print and web news. We investigate the locations and the countries involved in economic events (such as business or economic agreements), as well as the tone and the Goldstein scale of the news where the events are reported. We connect these elements to build three different social networks and to extract new network metrics, which prove their value in extending the predictive power of models only based on the inclusion of other economic or demographic indices. We find that the number of news, their tone, the network constraint of nations and their betweenness centrality oscillations are important predictors of the Gross Domestic Product per Capita and of the Business and Consumer Confidence indices.

Main finding: The number of news, their tone, the network constraint of nations and their betweenness centrality oscillations are important predictors of GDP per Capita and of BCI and CCI.

Elshendy, Mohammed	2017
Database	Events
IV/DV	IV
Variables	Economic events
Measure	Centrality and other measures from country networks of businesses interacting
Locations	Top 10 EU Countries
Analysis	Multi level regression
Period	2010-2016

The Power of the Street: Evidence from Egypt's Arab Spring

Acemoglu, Daron 2018

Review of Financial Studies */ Finance*

Abstract: Unprecedented street protests brought down Mubarak's government and ushered in an era of competition between three rival political groups in Egypt. Using daily variation in the number of protesters, we document that more intense protests are associated with lower stock market valuations for firms connected to the group currently in power relative to non-connected firms, but have no impact on the relative valuations of firms connected to rival groups. These results suggest that street protests serve as a partial check on political rent-seeking. General discontent expressed on Twitter predicts protests but has no direct effect on valuations.

Main finding: These results suggest that street protests serve as a partial check on political effect on valuations.

Database	Events
IV/DV	IV
Variables	Firm social sensitivity measure
Measure	Slope coefficient of regressing a firm's stock returns against a dummy variable that is one on the two trading days following any negative events
Locations	Egypt
Analysis	OLS regression
Period	2010-2013

Protesting in 'hard times': Evidence from a comparative analysis of Europe, 2000–2014

Quaranta, Mario

2016

Current Sociology / *Sociology*

Abstract: This article aims to demonstrate that during an economic crisis political protest increases. Recently, economic performance has suddenly worsened in Europe after a period of relative prosperity. The economic crisis affects citizens, who may turn to political protest to voice their discontent. However, the literature on social movements has often dismissed the link between economic performance and political protest, arguing that dissatisfaction is not sufficient for mobilization. Despite this argument, in the wake of the 'Great recession' that hit Europe, some scholars have argued that the economic crisis has been a factor in the mobilization of political protest. Nonetheless, a broad assessment of the link between the economic crisis and political protest has yet to be carried out. Based on a comparative longitudinal analysis of 25 European countries between 2000 and 2014, this article complements recent publications on the topic, and shows that economic performance, measured using both objective and subjective indicators, has a strong association with the number of political protests. The literature on the topic has not always presented conclusive results on this association. In contrast, this article provides updated and clear findings showing that the state of the economy matters for mobilization.

Main finding: Economic performance, measured using both objective and subjective indicators, has a strong association with the number of political protests.

Database	Events
IV/DV	DV
Variables	Unconventional political participation
Measure	Count of protest actions
Locations	EU Countries
Analysis	Linear Regression
Period	2000-2014

Funded by the Horizon 2020 Framework Programme of the European Union. Grant agreement number 822654

Predicting Social Unrest Events with Hidden Markov Models Using GDELT

Qiao, Fengcai

2017

Discrete Dynamics in Nature & Society / *Sociology*

Abstract: Proactive handling of social unrest events which are common happenings in both democracies and authoritarian regimes requires that the risk of upcoming social unrest events is continuously assessed. Most existing approaches comparatively pay little attention to considering the event development stages. In this paper, we use autocoded events dataset GDELT (Global Data on Events, Location, and Tone) to build a Hidden Markov Models (HMMs) based framework to predict indicators associated with country instability. The framework utilizes the temporal burst patterns in GDELT event streams to uncover the underlying event development mechanics and formulates the social unrest event prediction as a sequence classification problem based on Bayes decision. Extensive experiments with data from five countries in Southeast Asia demonstrate the effectiveness of this framework, which outperforms the logistic regression method by 7% to 27% and the baseline method 34% to 62% for various countries.

Main finding: This framework outperforms the logistic regression method by 7% to 27% and the baseline method 34% to 62% for various countries.

Database	Events
IV/DV	IV
Variables	Country social instability
Measure	GDELT event count
Locations	Cambodia, Malaysia, Indonesia, Philippines, Thailand
Analysis	Hidden Markov Models
Period	2001-2016

Funded by the Horizon 2020 Framework Programme of the European Union.
Grant agreement number 822654

Utilizing remote sensing and big data to quantify conflict intensity: The Arab Spring as a case study

Levin, Noam

2018

Applied Geography / *Geography*

Abstract: Tracking global and regional conflict zones requires spatially explicit information in near real-time. Here, we examined the potential of remote sensing time-series data (night lights) and big data (data mining of news events and Flickr photos) for monitoring and understanding crisis development and refugee flows. We used the recent Arab Spring as a case study, and examined temporal trends in monthly time series of variables which we hypothesized to indicate conflict intensity, covering all Arab countries. Both Flickr photos and night-time lights proved as sensitive indicators for loss of economic and human capital, and news items from the Global Data on Events, Location and Tone (GDELT) project on fight events were positively correlated with actual deaths from conflicts. We propose that big data and remote sensing datasets have potential to provide disaggregated and timely data on conflicts where official statistics are lacking, offering an effective approach for monitoring geopolitical and environmental changes on Earth.

Main finding: Some of the metrics quickly respond to fighting events (depending on their intensity, e.g., news reports, Flickr photos, night lights), others show lag times (e.g., numbers of asylum seekers).

Database	Events
IV/DV	IV
Variables	Violent events
Measure	Countries
Locations	19 MENA countries
Analysis	Regression
Period	2000-2016