institute of
development
studies

Working Paper
Volume 2023 Number 596

Big Data for Development Studies? An Innovative Methodology

Rajith W.D. Lakshman, Dolf J.H. te Lintelo, Hart Ford, Elissa Chattat and Madhubhashi Senanayake

October 2023

The Institute of Development Studies (IDS) delivers world-class research, learning and teaching that transforms the knowledge, action and leadership needed for more equitable and sustainable development globally.

Working Paper Volume 2023 Number 596
Big Data for Development Studies? An Innovative Methodology
Rajith W.D. Lakshman, Dolf J.H. te Lintelo, Hart Ford, Elissa Chattat and Madhubhashi Senanayake
October 2023

First published by the Institute of Development Studies in October 2023
ISSN: 2040-0209 ISBN: 978-1-80470-159-1
DOI: **10.19088/IDS.2023.053**

ORCID IDs: Rajith W.D. Lakshman **0000-0001-8178-9223**; Dolf J.H. te Lintelo **0000-0002-9216-498X**; Madhubhashi Senanayake **0000-0001-8706-4346**.

Suggested citation: Lakshman, R.W.D.; te Lintelo, D.J.H.; Ford, H.; Chattat, E. and Senanayake, M. (2023) *Big Data for Development Studies? An Innovative Methodology*, IDS Working Paper 596, Brighton: Institute of Development Studies, DOI: **10.19088/IDS.2023.053**

A catalogue record for this publication is available from the British Library.

The authors are thankful to the Dutch Research Council (NWO)-WOTRO for funding this work. The project is a collaboration between IDS, ACTED, and World Vision Lebanon.

Available from:
Institute of Development Studies, Library Road
Brighton, BN1 9RE, United Kingdom
+44 (0)1273 606261
ids.ac.uk

IDS is a charitable company limited by guarantee and registered in England
Charity Registration Number 306371
Charitable Company Number 877338

Working Paper
Volume 2023 Number 596

Big Data for Development Studies? An Innovative Methodology

Rajith W.D. Lakshman, Dolf J.H. te Lintelo, Hart Ford, Elissa Chattat and Madhubhashi Senanayake

October 2023

Big Data for Development Studies? An Innovative Methodology

Rajith W.D. Lakshman, Dolf J.H. te Lintelo, Hart Ford, Elissa Chattat and Madhubhashi Senanayake

October 2023

Summary

This paper makes a foundational methodological contribution to the analysis of big data for development studies. Big data has revolutionised the natural and applied sciences (and commerce). However, its use within development studies has been comparatively limited. This is despite a clear potential for devising innovative research methodologies to generate new academic knowledge, theory, and empirical insights of practical relevance to the broad field of international development and humanitarian policy, programming, and action. The paper uses the publicly available Global Database of Events, Language, and Tone (GDELT) as a source of near-real-time big data. The aim is to develop an academically robust methodology, that can be replicated and easily modified to suit the needs of a wide development studies scholarship, while also exploring its prospects for generating valuable actionable knowledge, through an established academic–practitioner partnership.

Keywords

Big data; Lebanon; GDELT; development studies; case studies; humanitarian action; media analysis.

Authors

Rajith W.D. Lakshman is a Research Fellow at the Institute of Development Studies, with over 15 years of experience in collaborative research on internally displaced persons (IDPs) and refugees. His work focuses on the intersection of gender, migration, wellbeing, and refugee studies. Rajith’s recent work examines the relationship between urbanization and forced displacement, specifically exploring the impact of IDP and refugee movements on host communities in the urban context. Rajith’s research sheds light on the importance of understanding the complexities of forced migration and the need for inclusive, community-led solutions.

Dolf J.H. te Lintelo is a Research Fellow at the Institute of Development Studies. He leads the IDS Cities Cluster. His research analyses the multi-scalar governance processes, actors, and state/humanitarian/development policies and practices that govern poor and displaced populations' incorporation into city spaces globally. He has an enduring interest in urban informality; food/nutrition insecurity, poverty, and wellbeing, and the ways in which marginal groups exercise (constrained) agency and contest structural factors of disadvantage. Dolf has also worked extensively on methods and metrics evaluating government policy and action on malnutrition.

Hart Ford is the Regional Director for the Middle East and North Africa for ACTED based in Beirut. Building on an educational background in socio-anthropology and economic development, Hart has been working across various aspects of humanitarian and development programming specifically within the Middle East and North Africa for the past 12 years. Her work has focused on tackling the issues born out of the interconnection between exclusion, poverty, and climate change across the levels of individuals, systems, and policies.

Elissa Chattat is the research and data analyst at World Vision Lebanon. Her research extends across various sectors including child protection, education, livelihood, food security, and WASH. With a background in psychology, Elissa has a longstanding interest in human behaviour and wellbeing and incorporates a human-centred perspective into her research. Beyond her research experience in psychology, she recently supported the development of a vulnerability map for Lebanon, alongside other research projects that focus on enhancing children's lives, education, and wellbeing.

Madhubhashi Senanayake is an experienced data scientist with a demonstrated history of working in the computer software industry. With a background in engineering, computer science, and research, Madhu has a strong interest in applications of natural language processing (NLP). In recent years, she has been involved in building a domain-agnostic framework for conversational AI for which she leverages machine learning and pretrained language models including large language models (LLMs) such as the infamous GPT. As an NLP enthusiast, Madhu enjoyed working on GDELT data for this project as it is built using NLP to understand worldwide news.

Executive Summary

This paper delves into the innovative potential of big data, with a focus on the publicly available Global Database of Events, Language, and Tone (GDELT), within the field of development studies. The rapid expansion of big data has transformed numerous sectors, yet its application within development studies remains limited. This study seeks to bridge this gap by developing a novel methodology for utilising GDELT data in a nuanced, case-specific manner.

The research unfolds against a backdrop of academic–practitioner collaboration, aiming to make the findings relevant to both academic scholars and development practitioners. Four case studies in Lebanon serve as the empirical basis, highlighting the use of event counts, event tones, and Global Knowledge Graph (GKG) themes to identify key moments within the contexts of the case studies. Notably, this approach is pioneering, offering valuable insights into media narratives on development topics.

However, the study grapples with challenges, primarily related to missing URLs within GDELT's data, hindering the confirmation of relevant events. This limitation necessitated a combination of approaches, including the use of secondary sources and keyword analysis of the URLs. Despite these challenges, the research underscores GDELT's potential in temporal and geographic mapping of events related to the cases and in discerning public opinion through media tone analysis. Additionally, the study suggests that GDELT's data can inform communication strategies and shape media narratives, particularly in the development sector.

In summary, this research contributes to the evolving field of big data in development studies, showcasing both the potential and challenges of utilising GDELT for nuanced case-specific analysis. It exemplifies the transformative power of big data within the development context, offering fresh perspectives and opportunities for research, practice, and policymaking in the ever-evolving landscape of international development and humanitarianism.

Contents

Figures

Tables

Acknowledgements

The authors are thankful to the Dutch Research Council (NWO)-WOTRO for funding this work and their support towards managing a challenging research journey. Several colleagues from ACTED and World Vision were part of the research but had not authored this paper. We wish to gratefully acknowledge them here: Elyan Farés, Ziad Harb, Anwar Mitri, Hady Naal, and Joelle Semaan. The authors are also grateful for the able project support from Rina Badran Lizarazo, Martin Gardner, and Rosa Lewis-Anthony.

Acronyms

API	Application Programming Interface
BD4D	Big Data for Development
CAMEO	Conflict and Mediation Event Observations
CSKC	Civil Society Knowledge Centre [Lebanon]
EPU	Economic Policy Uncertainty
FC	Fuel Crisis
GC	Garbage Crisis
GCAM	Global Content Analysis Measures
GDELT	Global Database of Events, Language, and Tone
GKG	Global Knowledge Graph
GoL	Government of Lebanon
GPS	Global Positioning System
GPT	Generative Pre-training Transformer
HRW	Human Rights Watch
IDP	internally displaced person
INGO	international non-governmental organisation
LLM	large language model
NGO	non-governmental organisation
NLP	natural language processing
OR	October Revolution
RR	Refugee Returns
SQL	Structured Query Language
TB	terabyte
URL	Uniform Resource Locator
WASH	water, sanitation, and hygiene
WV	World Vision

1. Introduction

The phenomenon of big data, characterised by remarkable expansion in data volume, velocity, and variety (Laney 2001, as referenced in Diebold 2021) has permeated every facet of modern society. This proliferation is in line with the predictions of those who were excited about this possibility a decade or so ago for commerce in particular (Chen, Chiang and Storey 2012; LaValle *et al*. 2011) and more broadly (Manyika *et al*. 2011; Mayer-Schönberger and Cukier 2013). Diverse disciplines such as astronomy, town planning, cancer research, climate change, and so forth have transformed due to the advent of big data (Cukier and Mayer-Schoenberger 2013; Kitchin 2022). For instance, real-time data on traffic conditions has made adaptive traffic management systems feasible in urban environments (Kitchin 2022). In the field of medicine, big data has revolutionised medical administration, enabled early detection of health issues, and facilitated personalised patient management, among other advancements (*ibid*.). However, despite the profound impact big data has had in numerous other fields, its application within development studies has been relatively limited. This paper seeks to address this gap by exploring the potential for an innovative big data-inspired research methodology to generate empirically relevant insights in the expansive domain of international development and humanitarian policy, programming, and action.

The limited use of big data in development studies is evident; out of 177,541 matches for a simple search for 'big data' on *Web of Science*, only 251 are from development studies.[1] This limitation is not because big data is irrelevant for development. On the contrary, Hilbert (2016), reviewing a range of studies, suggests that big data applications have already transformed critical development areas such as health care, economic productivity, and security. Advances in automated data collection, telecommunication networks, data storage, and computing capacity/speed over the last two decades was a key factor in this (*ibid*.). However, such technological innovations are unevenly distributed, with developing countries lagging far behind the developed (Andrejevic 2014; Boyd and Crawford 2012; Hilbert 2016). Boyd and Crawford (2012) describe this as a new kind of 'digital divide' and Hilbert (2016) links developing country problems in this regard to structural issues such as inadequate computing infrastructure, limited human capital, a shortage of economic resources, and underdeveloped institutional frameworks (Hilbert 2016: 164). With weak dissemination of big data in developing countries, it is not surprising that development studies is a comparatively late arrival in the big data

[1] *The Journal of Development Studies* carried only six articles matching this simple search; *World Development* had two. These search results are dated 15 September 2023.

scene; after all, development studies has long concerned itself predominantly with developing countries (Sumner 2022).

Despite the challenges outlined above, big data dissemination has begun in developing countries. Kshetri, Fredriksson and Torres (2017) document a range of examples of big data applications in developing countries, even if only in a limited number of sectors. One of these examples, Aadhaar, the national population register in India, is already the world's largest biometric database with nearly 1.4bn Indian residents registered.[2] The wider availability of cloud computing in place of prohibitively expensive hardware for big data analyses, and the role of multinationals as technology diffusers seem important drivers in the initial stages of big data uptake in developing countries (Kshetri *et al*. 2017; Mann 2018). Open data initiatives by governments and the public sector, and data-sharing initiatives by global tech giants opens further opportunities for big data use in the developing world (Buckee, Balsari and Schroeder 2022). The emergence of humanitarian actors as big data users in developing countries also provides a boost to the dissemination of big data know-how within these contexts (Kondraganti, Narayanamurthy and Sharifi 2022; Sharma and Joshi 2020).

While all of the above initiatives may help big data diffusion in developing countries, scientists in critical data studies (Boyd and Crawford 2012) and those working on the political economy of data (Mann 2018) warn about the threats big data would entail for those in developing countries. The threat to the personal freedom of big data subjects and at a broader level to democracy is indeed a concern (Andrejevic 2014; Cieslik and Margócsy 2022; Couldry and Mejias 2020). 'Datafication', the process which leads to big data through the quantification of human life using digital information is at the centre of these concerns (Mejias and Couldry 2019). Even development and humanitarian agencies are complicit in deploying population surveillance-based solutions that pose inadvertent yet serious threats to individual rights (*ibid*.). This is particularly true in 'contexts where people, laws, and human rights are the most fragile' (Milan and Treré 2019). Due to the extraction of value from data subjects within a context of unequal power dynamics, some scholars describe big data transformation as 'data colonialism' (Couldry and Mejias 2020). In addition to these overarching concerns, there are others that are more methodological in nature. The issue of sampling, for example, where it is argued that big data gives access to the population is just not right: 'It is an error to assume "people" and "Twitter users" are synonymous' (Boyd and Crawford 2012: 669). This is a serious concern in developing countries where digital penetration is poorest. There is a significant risk that big data in such countries could be biased (Hilbert 2016).

[2] See **https://uidai.gov.in/aadhaar_dashboard/index.php**.

Despite these valid critiques, there remains value in interrogating big data in development studies, provided this is done cognisant of these concerns. For instance, Hilbert (2016) employs his Big Data for Development (BD4D) framework to identify applications of different types of big data analysis and explore their potential for enhancing development outcomes. The main advantage of big data applications in research is that their results will be available for free and in real time; whereas if the same research is done through traditional methods, it will cost much more and take a lot longer (*ibid*.). Hilbert considers word tracking applications using sources such as social media, blogs, news, web pages, and so forth for unemployment estimation and for predictions such as flu pandemics. Similarly, tracking of locations via mobile signals and GPS tracking has been used for traffic monitoring, crime detection, and so on. Tracking of nature, for example, in remote sensing of weather patterns, tracking of individual transactions through the analysis of sales transitions, and tracking of production, especially of resource-extracting entities, also offers a range of opportunities for development actors and academics alike (Hilbert 2016).

The purpose of the paper is for a team of researchers, consisting of academics and practitioners in development, to jointly explore big data sources and co-develop a methodology for mainstreaming big data use in development studies. The specific big data set used is the Global Database of Events, Language, and Tone (GDELT), the largest publicly available global media data set (Yonamine 2013). GDELT data has certainly been used in development-related research (Coniglio, Peragine and Vurchio 2023; Czvetkó *et al*. 2021; Honti *et al*. 2021; Levin, Ali and Crandall 2018; Qiao *et al*. 2017). Across all of this extant body of work, as far as we know, GDELT is used to assess/interrogate the ‘big picture’. For example, Coniglio *et al*. (2023) look at all of Africa, Czvetkó *et al*. (2021) at the whole world, and so on. In this paper, on the other hand, we focus on using GDELT as a more nuanced, disaggregated, case-specific data source. This novel focus is expected to open up a much wider set of big data applications in line with the need in development studies to understand contextually specific development dynamics, that acknowledge difference, rather than just heavily aggregated big picture analysis that erases difference.

The paper is organised thus. The next section develops the methodological approach of the paper which is based on four case studies of how a team of researchers from academic and practitioner backgrounds collaboratively made sense of big data analysis based on GDELT data. This will be followed by the four case studies highlighting what we learned about the cases by engaging with GDELT. The case studies will be followed by the final section on discussion and conclusion.

2. Methodology of the media-based analysis of cases

This paper, as highlighted earlier, is motivated by the comparatively limited uptake of big data-inspired methods in development studies, despite its potential. We, as a team, already had firsthand experience of some of the difficulties which contribute to the state of big data uptake within the field (Lakshman *et al*. 2020). The team represents two academics (the Institute of Development Studies), two practitioners (ACTED and World Vision in Lebanon), and a data scientist. The in-country research partners who are from practitioner backgrounds are ideally positioned to check whether GDELT findings were consistent with the ground truth and help with the interpretation of findings, and with considering how such findings and approaches may have potential for organisational practice, programming, and action.

2.1 The case study method

A large part of why we focus on Lebanon is to do with the present work being a continuation of the research team's previous work. Yet the country is also eminently suitable for a study on big data inquiry for development studies. Particularly as the data we use (GDELT version 2.0) is available only after 2015, and the post-2015 history of Lebanon is tumultuous, to say the least. The four cases are *a priori* important episodes/periods in the real-world development context defined thus: an instance of collaborative use of GDELT data by academics and practitioners to illustrate possible opportunities and costs of big data analysis in development studies.

Looking at these cases in more detail is expected to shine a light on what is stopping others in development studies from using big data in their research and practice. The selection of the four case studies was led by ACTED and World Vision using the selection guidelines in Annexe C. This was a two-step process where initially a shortlist of eight case studies was selected, from which a final four were selected in the second step. Table 2.1 summarises the four cases. As noted above, all four were of specific interest to development/humanitarian practitioners. The cases cover a mix of older and newer cases. This mix was important because the age of the case was a big factor in the availability of source news material.

Table 2.1 Summary descriptions of the cases

Long name	Short name	Year	Description
Garbage Crisis	GC	2015	A series of agitations/protests related to the garbage disposal problem which morphed into a bigger movement of civil demonstrations, e.g. the ‘You Stink’ movement.
Refugee Returns	RR	2018	Politically driven initiatives for the return of refugees back to Syria.
October Revolution	OR	2019	The outbreak of coordinated protests linked to the accumulated crises over the preceding weeks.
Fuel Crisis	FC	2021	Severe fuel shortages experienced over several months. This is connected to fuel subsidies/prices and a series of power outages in the country.

Source: Authors’ own.

The approach here is to use three data sets from the GDELT collection (the GDELT Events dataset, Mentions data set, and Global Knowledge Graph, GKG, data set) and develop a robust methodology to access, analyse, and interpret information gleaned from those sources. The case study approach focuses on specific ways of employing GDELT data to investigate how the media has depicted these significant cases from post-2015 Lebanon. It is hoped that both academics and practitioners in development fields will gain valuable insights into the portrayal of these events and their implications.

2.2 Access and analysis of GDELT data

There are three main sources of accessing the GDELT data (Williams 2020). The first two, (1) the full-text Application Programming Interface (API) and (2) the raw data files, are available from the website of the GDELT project.[3] The API approach, while user-friendly, works only for a window of three months. Clearly, it will not work for the current purpose, where we are keen to look at historical data up to 2015. The use of raw data files running into multiple terabytes (TBs) is not possible within the present project because of storage and computing

[3] These can be accessed via **https://analysis.gdeltproject.org/** and **https://www.gdeltproject.org/data.html#rawdatafiles** respectively.

limitations. So we opted for the third option, accessing the GDELT data via the Google BigQuery platform, which contains the full array of GDELT data sets including the three used here. We used a paid Google BigQuery account for this work.[4]

Structured Query Language (SQL) was used for filtering and pre-processing the data within the BigQuery platform before downloading it for further analysis. Annexe B includes all of these queries with some annotation. The time filtering, country filtering, thematic filtering, and so forth were all done on the Google BigQuery platform as they all required processing of data at a terabyte scale. These filtering stages reduced the volume of data to a size that is manageable locally on our personal computers.[5]

The filtering of the GDELT data sets involved the use of **Event codes** and **GKG themes** to first identify all events in Lebanon during a calendar year relevant to the case (see Table 2.1). The codes were then used to further refine the filtering to capture events that make specific references to the cases we are interested in.[6] Though quantitative filtering using SQL on the Google BigQuery platform was central to this, the process also involved qualitative checking to ensure that the filtered news articles were relevant. Therefore, this may be described as an iterative mixed-method approach. During each iteration, quantitative filtering was followed by qualitative checking of the resulting news reports for relevance to the case study.

2.3 Filtering GDELT: from big data to small data

As noted above, keywords/theme selection is a key element in the quantitative filtering of relevant events. Our purpose here was to be able to filter events related to specific cases from out of tens of thousands of events coded for Lebanon each week. The GDELT uses the predefined Conflict and Mediation Event Observations (CAMEO) coding manual to machine-code the events within its events database. This CAMEO categorisation of media events can be a useful tool in research and practice (Lakshman *et al*. 2020). Yet the conflict focus of the CAMEO coding is found to be too restrictive for the purpose of the present research. For example, the CAMEO code for protests (GDELT Event root code 19) probably captures all of the events around Lebanon’s garbage crisis in 2015, our first case study. But the problem is that it would also capture a large number of other protest events that would not refer to garbage or solid waste problems.

[4] While Google allowed for some BigQuery analysis free of charge, the free quota was not sufficient for this research. For this research, including various experimental steps, we paid less than £150.

[5] Computer resources are a key impediment to the use and analysis of big data, particularly for the development studies community.

[6] Annexe A describes these data-wrangling stages in steps 1 and 2.

We clearly needed to use a more nuanced schema of codes. This is exactly what is possible with the GKG themes. Going back to the garbage crisis example, the GKG theme 'WB_1797_SOLID_WASTE' was a better fit for what we wanted to capture in relation to that case.

The GKG data set records individual news articles, extracting detailed information on a range of themes mentioned in an article in addition to other information. The complete list of GKG themes is constantly expanding.[7] It included an expanding list of themes developed by GDELT, as well as several externally curated lists of themes such as those based on the topical taxonomy of the World Bank Group (World Bank 2016; The GDELT Project 2015) and on the Economic Policy Uncertainty (EPU) index by Baker, Bloom and Davis (2016). These themes, as explained above, can be used to filter GDELT events so that the results better capture events that are related to the case studies (GDELT 2021).

2.4 Aggregation and analysis of Event counts and tones

These data-wrangling stages were completed on the Google BigQuery cloud platform as the relevant GDELT data sets, even after they were filtered, were either beyond our local storage capacities or would take too much time to download. The outputs of the data-wrangling steps, however, were more appropriate for our storage and bandwidth capacities. We downloaded three data sets: (1) a weekly data set with event counts and average tone, (2) geolocation of the filtered events at a subnational location,[8] and (3) a data set with the URLs of matching news reports. We used two approaches to download these from Google BigQuery to our local computers depending on how we did the processing thereafter. The data that was further analysed in the R program[9] was directly downloaded to R using **bigrquery** package,[10] and the data that was further processed using Google App Scripts was directly saved as a Google Sheet.

The case studies in the next section of the paper bring together quantitative data and qualitative findings from manual scrutiny of available news reports. The

[7] At the time of writing in September 2023, there were 60,773 themes across the full global GKG data set. The corresponding figure in November 2021 was 59,314 (GDELT 2021). The GKG theme count for Lebanon for each of the cases we had looked at was less than 10,000.

[8] Most matching events only have country-level Global Positioning System (GPS) identifiers (centroid for Lebanon as a whole) which was not useful for the purpose of understanding the in-county geographic spread of events. These media events were excluded when preparing the maps. The maps, therefore, do not capture all matching events but only those that mention in-country locations.

[9] See **https://www.r-project.org/**.

[10] See **https://cran.r-project.org/web/packages/bigrquery/**.

quantitative analysis is presented as two sets of time series data (weekly counts of matching events and the weekly average tone of these events) and maps of the counts. GDELT defines 'tone' as the collective sentiment expressed across all documents that reference a specific event during its initial appearance in a 15-minute update.[11] GDELT uses the Global Content Analysis Measures (GCAM) system that runs each document through an array of content analysis systems to assess several thousand latent dimensions for each article.[12]

Time-series patterns of the weekly counts of events and time-series patterns of the weekly average tone of the events were used to identify important points in the stories around the case studies. The qualitative interrogation was based on lists of URLs for news filtered as relevant to the case. A large proportion of the URLs for the matching events/news were not live at the time of this research.[13] The manual checking of available news reports is expected to contribute to the understanding of the events related to the case studies in general but also to the understanding of what happened during important points identified through time-series analysis.

[11] This sentiment is quantified as an average value, falling within a numerical range spanning from -100 to +100. However, the majority of articles typically exhibit sentiments between -10 and +10.

[12] **https://blog.gdeltproject.org/introducing-the-global-content-analysis-measures-gcam/** gives more details about the content analysis systems used.

[13] This proportion increased with the older case studies.

3. Case study analysis

In this section, we engage with the GDELT data for selected cases with the aim of identifying whether the three GDELT data sets considered here are helpful in developing an evidence-supported, reasonably nuanced picture of the cases. For each case, the time-series analysis and the mapping of the quantitative data are available in a single graph. Qualitative reflections from having consulted available news from media URLs available in GDELT are also included in the case studies. When such news reports are cited, the relevant bibliography item will also include GDELT's unique event ID.[14]

3.1 Case 1: Garbage Crisis of 2015

Waste management has always been an issue in the recent history of Lebanon. After its privatisation in 1994 following the civil war, waste management became highly influenced by political corruption (Khneisser 2019). The piling up of garbage on the streets in the summer of 2015, this case study, brought on by the closure of the Naameh landfill also has corruption undertones (*ibid*.). Despite anticipating this problem, political leaders couldn't agree on a new arrangement to avoid this crisis (*ibid*.). For this case study we are only interested in events around the Garbage Crisis of 2015 (GC hereafter).

As explained in the methodology section above, we first filtered the GDELT Events database for all events in Lebanon from 2015. These were then merged with the GKG data set so we could use GKG themes to further filter the events. This is important because we wanted to use the World Bank Topical Taxonomy available in the GKG data for this. The particular approach was to filter the events that were coded under the topic 'WB_1797_SOLID_WASTE' (WB1797 hereafter).[15] This code helped us filter waste-related topics, which corresponded closely with GC-related events. The filtered data was then downloaded which included the weekly summarised data, geolocations, and the URLs.

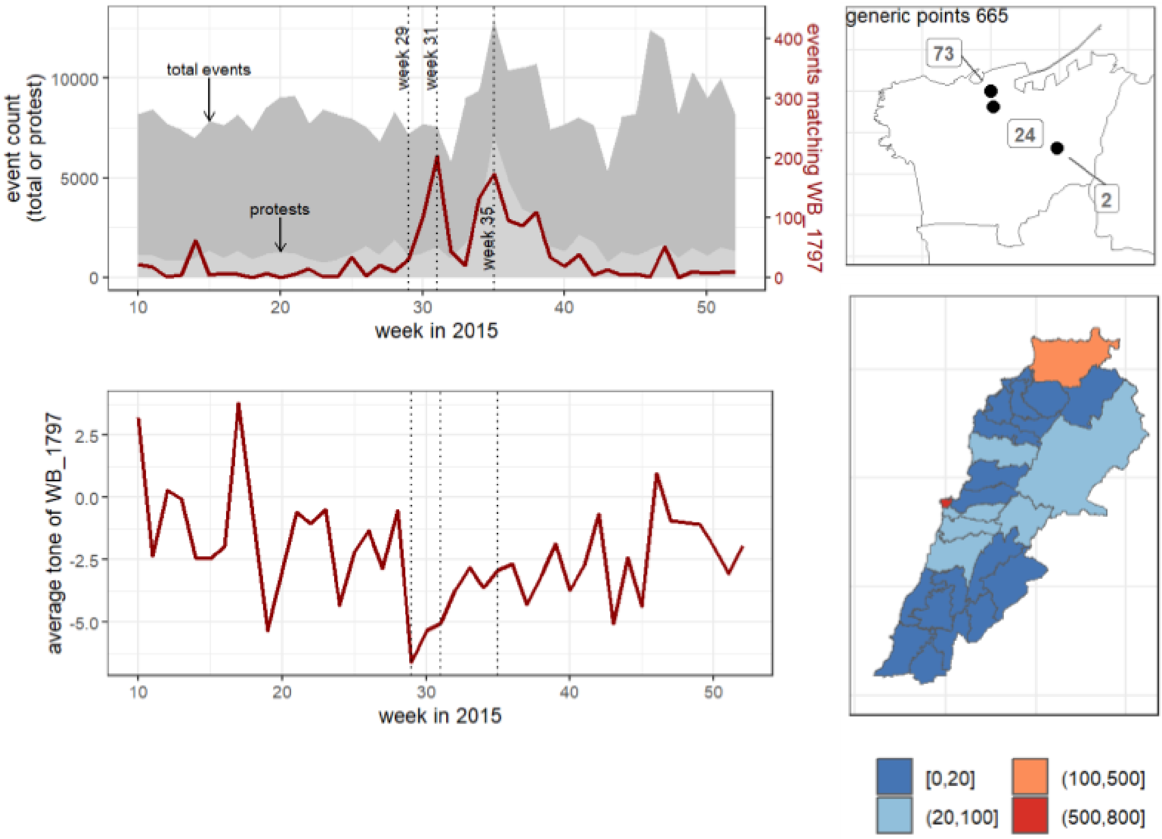
The dark red line in Figure 3.1 plots the weekly counts (top graph) and the weekly average tone (bottom graph) of the WB1797 events. It is clear that the selected events were only a very small fraction of the total events counts for the period (dark grey area graph) or of total protest-related events (light grey area graph). This data shows that the protests culminated in week 35 which included the dates of mass protest movements in late August 2015. The regional and the

[14] A single news report can include multiple unique GDELT events. So the GDELT news items we cite here can have information about unique events other than the one for which the event ID is cited in the bibliography. Here is an example of how a unique event ID is referenced in the bibliography: '[#880715130]'.

[15] See **https://vocabulary.worldbank.org/taxonomy/1797.html**.

Beirut maps show the distribution of places where the matching WB1797 events happened. Most of the geocoded media events that were matched were from Beirut. Most of them (2021 events) were generically geocoded to the centroid of the district; the rest can be traced as plotted in the map of Beirut in Figure 3.1. The average tone of WB1797 in the second graph in Figure 3.1 suggests that week 29 was a pivotal juncture in GC: the week coincides with the shutdown of the Naameh landfill on 17 July (Naharnet Newsdesk 2015a, 2015b). The event count data suggests that media interest in solid waste increased rapidly until week 31, followed by a lull and then another boost leading to a peak in week 35.

Figure 3.1 GDELT Events and GKG data for the 2015 garbage crisis in Lebanon[16]



Source: Authors' own.

In what follows, we zoom into the information about filtered WB1797 events to learn more about GC and how it evolved. To do that, we downloaded the URLs of WB1797 events for 2015 and examined them for relevance to GC.

[16] GDELT version 2 on which this research was based was only fully operational after ten weeks into 2015.

Even though GDELT data for Lebanon in 2015 had 1,430 matching events, each of those events was published in multiple news sources coded with WB1797, which meant that the URL count for the year was 2,271. Many of these URLs from back in 2015 were not live at the time of this research. We used a tool developed using Google's URL Fetch Service (see Annexe A) to resolve most of these, so the researcher could easily focus on URLs that have a high chance of connecting them to a new item that they could use.

The important points in the time series of weekly WB1797 counts can be confirmed through specific URLs in the database. For example, week 29, the lowest point in tone, can be identified in the news reports as when the landfill was closed (Al Araby 2015; Naharnet Newsdesk 2015b). The peak in WB1797 counts in week 31 reflects a mix of GC-relevant issues around attempts to solve the crisis including temporary ones and reporting of signs of further escalation of the problem (Al Bawaba 2015; Al-Manar 2015). Week 31, from 27 July until 2 August 2015, yielded a number of relevant news reports. Specifically, 28 July marks the day a movement referred to as 'You Stink' organised its first spontaneous demonstration (CSKC 2016; Khneisser 2020). However, the URLs for week 31 which were still working at the time of the research did not mention the movement at all. The URLs in subsequent weeks certainly did, in a very big way. It seems that week 31 is not just a random peak but one that points to #YouStink movement even though traces of it may be hard to find in GDELT now after eight years.

The peak in week 35 also compares with evidence around weeks 29 and 31 in terms of results. The peak on week 35 for WB1797 contrasts against the much larger peak for 'protests' (the light grey area graph) during the same week. It suggests that the energy of the protests which started with the focus on waste/garbage had by then (week 35) spilt over and galvanised a wider protest movement which culminated in the massive protests in August. Khneisser (2019: 1113) describes these events thus:

> Security forces' unexpected crack-down on demonstrators on August 19th presented a turning point for the protest movement, compelling larger crowds to the streets in the next demonstration on August 22nd. Clashes increased on August 23rd and 24th between the thousands of protestors and the riot police, leaving many injured and one dead.

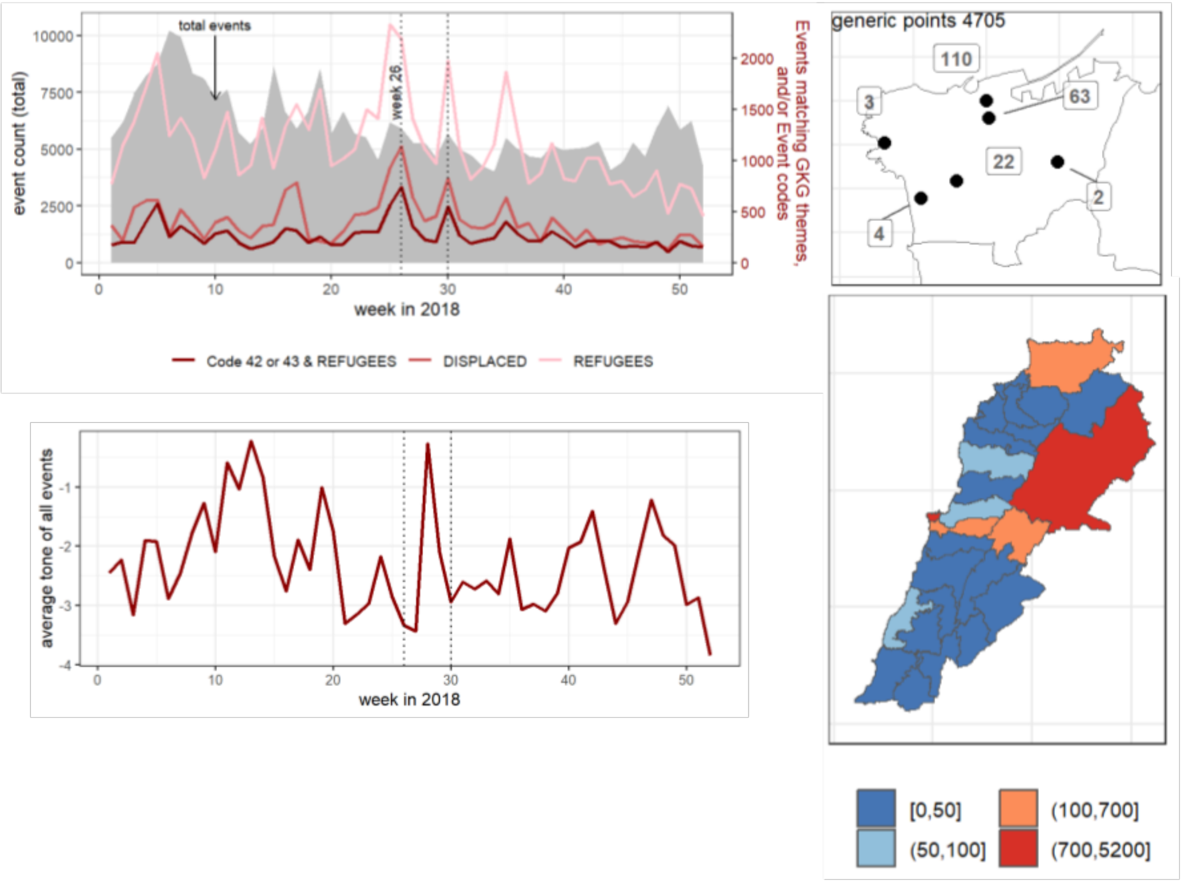
In addition to news directly related to GC, the URLs revealed news articles about events that happened around the GC theme. These provide grounds for comprehensive research work that goes beyond the event of interest itself and expands to incorporate related events that surround it.

3.2 Case 2: Refugee Returns

By 2018, Syria's prolonged conflict had entered its eighth year, resulting in one of the most catastrophic humanitarian crises of the twenty-first century. Millions of Syrians had fled their homes, seeking refuge in neighbouring countries, with Lebanon hosting over 1 million Syrian refugees at its peak (GoL and United Nations 2018). Given the significant demographic burden, Syrian refugee presence weighed heavily on Lebanon's already fragile economic, social, and political landscape (*ibid*). Lebanon, a country with its own civil strife, found its infrastructure, economy, and social services under enormous strain due to the influx of refugees (te Lintelo *et al*. 2018). The demographic change stoked fears among various Lebanese factions that the predominantly Sunni Syrian refugee populace could alter Lebanon's political fabric (Fakhoury and Ozkul 2019; Geha and Talhouk 2018). Economic pressures led to heightened tensions between Lebanese locals and Syrian refugees, often leading to accusations against Syrians for the country's rising unemployment, decreased wages, and strained public services (Turner 2015). Political discourse frequently framed the refugee presence as a 'burden' or 'threat', with some parties capitalising on these sentiments to advance their political agendas.

Moreover, as certain areas in Syria began to witness a relative decline in hostilities in 2018, there were increased calls, both domestically within Lebanon and from the Syrian government, for refugees to return (HRW 2019). For Lebanese politicians, encouraging returns might appease local constituencies and alleviate economic strains. Yet many refugees were apprehensive about the voluntary nature of these returns, fearing persecution, conscription, and further displacement in Syria. Reports of detentions and forced conscriptions of returnees further exacerbated these anxieties (*ibid*.). The safety and voluntariness of returns thus became central themes for human rights organisations and international observers. There were also several push factors at play (UNHCR 2019).

Figure 3.2 GDELT Events and GKG data for 2018 events relevant to the Refugee Returns issue



Source: Authors' own.

The return of some Syrian refugees during 2018 (RR hereafter) happened against this background and attracted much media attention. It should mean that a media source such as the GDELT should document a story like RR in some detail. However, because the country is replete with many things to do with refugees, filtering media events related to only RR was difficult. For example, the events coded under the GKG theme 'refugees' (pink (or lightest) line in Figure 3.2) may have captured the RR events, but it will also capture a host of other events connected to refugees in Lebanon. Therefore, we combined the 'refugees' theme in GKG database with event code 42 (make a visit) and event code 43 (host a visit) in the Events database which permitted us to identify RR-relevant events in GDELT.[17] The count of these filtered events, as well as their average weekly tone, is graphed in dark red in Figure 3.2. The maps show that a

[17] CAMEO 042, make a visit, is described as 'Travel to another location for a meeting or other event' and CAMEO 043, host a visit, as 'Host or receive a visitor at residence, office or home country' (Schrodt 2012: 29).

large majority of events matching our criteria (event code 42/43 + refugees) can be traced to Beirut (4,909 events) but also to Baalbek (944 events). The quantitative analysis above strongly suggests that a few weeks of build-up of relevant events in June 2018 culminated in a peak in week 26. This was followed by a relative lull and a smaller peak in week 30. In what follows, we examine the underlying news articles from weeks 26 (from 25 June to 1 July) and 30 (from 23 July to 29 July).

In terms of the geographic clustering of events, the majority of these state-coordinated initiatives for the return of Syrian refugees from Lebanon in 2018 were focused on specific regions in Lebanon where large concentrations of refugees were present. The Lebanese border towns and areas in the Bekaa Valley, close to the Syrian border, were significant points of departure for these organised returns. Baalbek was a hotspot for such activities, with towns such as Arsal witnessing multiple rounds of such returns (AFP 2018; Naharnet Newsdesk 2018). While the border areas and the Bekaa Valley witnessed a more significant number of these operations due to their proximity to Syria, Beirut, housing a significant number of Syrian refugees, was not excluded. The Lebanese General Security organised buses from various points in Lebanon, including Beirut, to facilitate the repatriation of those willing to return to Syria. The large majority of data points related to Beirut are thus more likely due to political statements from political and military officials or UN and other aid agencies responsible for refugee movements based out of the country's capital, noting that disagreements arose surrounding the voluntary nature of these returns and whether or not they constituted refoulement.

In terms of the peak time period of related events, the majority of state-facilitated returns of Syrian refugees from Lebanon to Syria intensified during the second half of the year, particularly from July onwards. This push was in line with the increasing rhetoric from Lebanese politicians and military officials regarding the need for refugees to return, especially given the decreasing hostilities in certain parts of Syria. It is important to note that a key peak was recorded in week 26 around a particularly notable coordinated initiative reported by the Syrian government as a voluntary return programme (Gamal-Gabriel 2018), involving the return of hundreds of Syrian refugees (472 according to one political official quoted by Mroue 2018) from Aarsal (Saad 2018). In week 30, there were other significant return incidents, for example, from Aarsal through the Zamrani border crossing on 23 July, with different news sources citing between 800 and 1,000 individuals returning (AFP 2018) and the announcement of an expected return of 1,200 individuals from Masnaa on 28 July (Hatoum 2018). However, it is also important to note that within the GDELT database, a number of events in these peaks do not correspond to this event, even within the same event code. When taking the GDELT database and isolating the active links, we can see that the date with the most active links is 28 July. While this date corresponds to the announcement of the Masnaa return mentioned above, it also corresponds to other news announcements concerning a meeting between Lebanese and Russian officials to agree on a coordinated refugee return plan, which was dubbed the 'Russian Initiative' (Obeid 2018; Rose 2018).

3.3 Case 3: October Revolution

The Lebanese Revolution, often referred to as the 'October Revolution' (OR hereafter) began on 17 October 2019. This civil uprising was not an isolated event but the result of decades of socio-political challenges, systemic corruption, and economic distress. While the immediate trigger was tied to new tax proposals, the underpinnings of the revolution had roots in long-standing issues related to governance, economic distress, and a complex sectarian political system (Khatib 2022; Yahya 2019).

Rooted in the 15-year civil war that ended in 1990, Lebanon's political structure relies on a sectarian power-sharing system established by the Taif Agreement.[18] However, this structure institutionalised sectarian divisions and facilitated clientelism, with political leaders offering patronage in return for loyalty, leading to widespread corruption and inefficiency in the country (Makdisi and Marktanner 2009). Transparency International's Corruption Perceptions Index frequently ranked Lebanon low, indicating high levels of corruption (Transparency International 2019). The combination of clientelism and corruption meant that the needs and demands of ordinary Lebanese were often overlooked, leading to growing public disillusionment with the political elite.

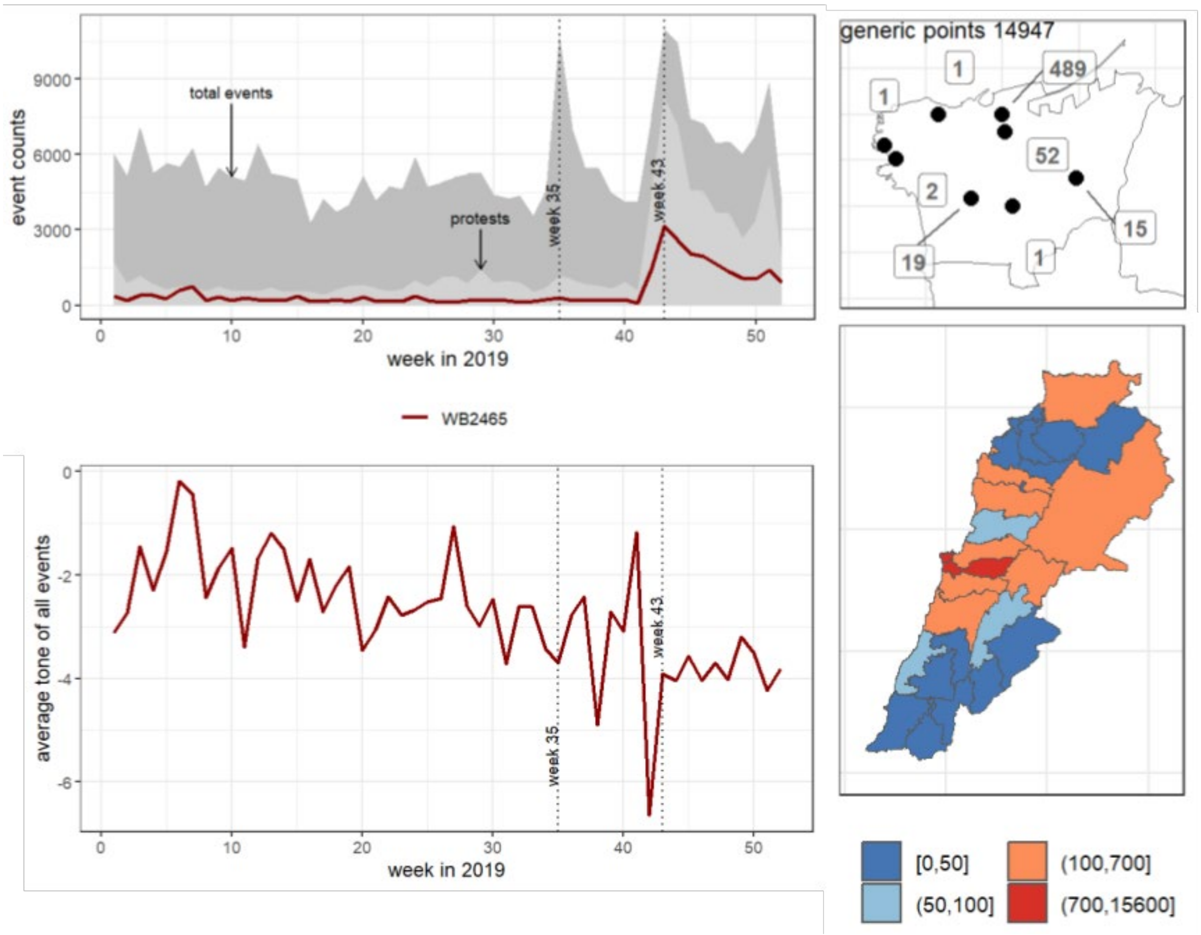
In addition to this complex political and governance backdrop, Lebanon was grappling with a looming economic collapse. With one of the highest public debt ratios worldwide, the country faced dwindling foreign reserves, stagnating growth, and rising unemployment (World Bank 2020). Compounded by the effects of the Syrian civil war on its economy and a refugee crisis, Lebanon's infrastructure was strained, and public services were under-resourced. Chronic mismanagement resulted in regular power cuts, unclean drinking water, and unreliable garbage collection. Significant protests related to waste management in 2015 had not resulted in more accountable governmental services and had already begun to forge momentum amongst citizens across party lines for reform (Yahya 2020).

In October 2019, just days before OR, wildfires spread across Lebanon's Chouf and Metn regions (Al-Manar TV Lebanon 2019). The government's inadequate response to these wildfires further exposed its inefficiencies and corruption (Chehab 2019). Lebanon's firefighting helicopters were grounded due to a lack of maintenance, compelling the government to seek foreign assistance in

[18] See https://peacemaker.un.org/lebanon-taifaccords89.

combating the fires (Abumaria 2019). During the same period, the government proposed new taxes on gasoline, tobacco, and even internet voice calls via apps such as WhatsApp (Al Jazeera 2019). These proposals were seen as out of touch, especially in the context of the economic downturn and public perceptions of elite corruption, leading to spontaneous demonstrations (Francis and Kanaan 2019).

Figure 3.3 Events and GKG data related to the October Revolution 2019



Source: Authors’ own.

OR is hard to miss in the time series graphs in Figure 3.3. Week 42 which includes 17 October 2019 is the starting point of the flurry of events peaking in week 43 (11,021 events) as illustrated in the graph of total event counts (dark grey area graph). However, the graph is also a good example of the confounding factors researchers and practitioners encounter when using this data: (1) the peak in week 35 seems equally important with a nearly identical total events count and (2) the events data series for Lebanon in 2019 has an average weekly count of about 5,000 events which suggests that we should expect a similar number of events in week 43, for example, to not be directly related to OR. So we needed a robust way to filter out the peaks such as the one in week 35, which corresponded with the Israel drone attacks in August 2019, and to filter out irrelevant events from the relevant window for OR. After iteratively experimenting with several World Bank thematic topics, we decided 'WB_2465_REVOLUTIONARY_VIOLENCE' (WB2465 hereafter) as the one that best captured news items with direct reference to OR while effectively filtering out the twofold confounding events noted above.[19]

In Figure 3.3, the weekly counts and the average weekly tone of WB2465 events are plotted in dark red line-graphs. As with the total events count, the counts of WB2465 also recorded a massive jump in week 42 (from 105 events in week 41 to 1,422 in week 42, before peaking at 3,153 in week 43). The sharp decline of the average tone of WB2465 events, from -1.1 in week 41 to -6.6 in week 42, also says a lot about the way things deteriorated quite quickly before the average tone stabilised around the -4 level during the rest of 2019. The country/Beirut maps plot the locations of events that were geolocated at the subnational level. Beirut (15,527 events) followed by Baabda (1,785 events) had the highest counts of subnationally geocoded WB2465 events. The map of Beirut identifies possible hotspots of OR.

It is notable from the clustering of the events within the Beirut district that 489 events were recorded within a single location, which corresponds to Martyrs' Square, a historically significant site where protests, gatherings, and national celebrations traditionally take place. Moreover, during the civil war, Beirut was divided into East and West, with a demarcation line separating the predominantly Muslim sectors from the Christian ones extending towards Martyrs' Square. Post-civil war, the square became a symbol of unity, reflecting the country's desire to move past sectarian divisions. This square became the epicentre of OR, where protests, gatherings, intellectual/political discussions, cultural events, and media coverage converged.

In what follows, we examine the news items from weeks 42 (14 October to 20 October) and 43 (21 October to 27 October) and use them to describe what happened during those two weeks in relation to OR. Week 42 corresponds to a number of the key events that led up to OR, such as the wildfires mentioned above in Metn and Chouf on 14 and 15 October, as well as the mass protests which were kicked off on 17 October in response to the announced austerity measures including the WhatsApp Tax.

In the subsequent days, extending into week 43 and beyond, the demonstrations grew exponentially, with hundreds of thousands taking to the streets across the

[19] See **https://vocabulary.worldbank.org/taxonomy/2465.html**.

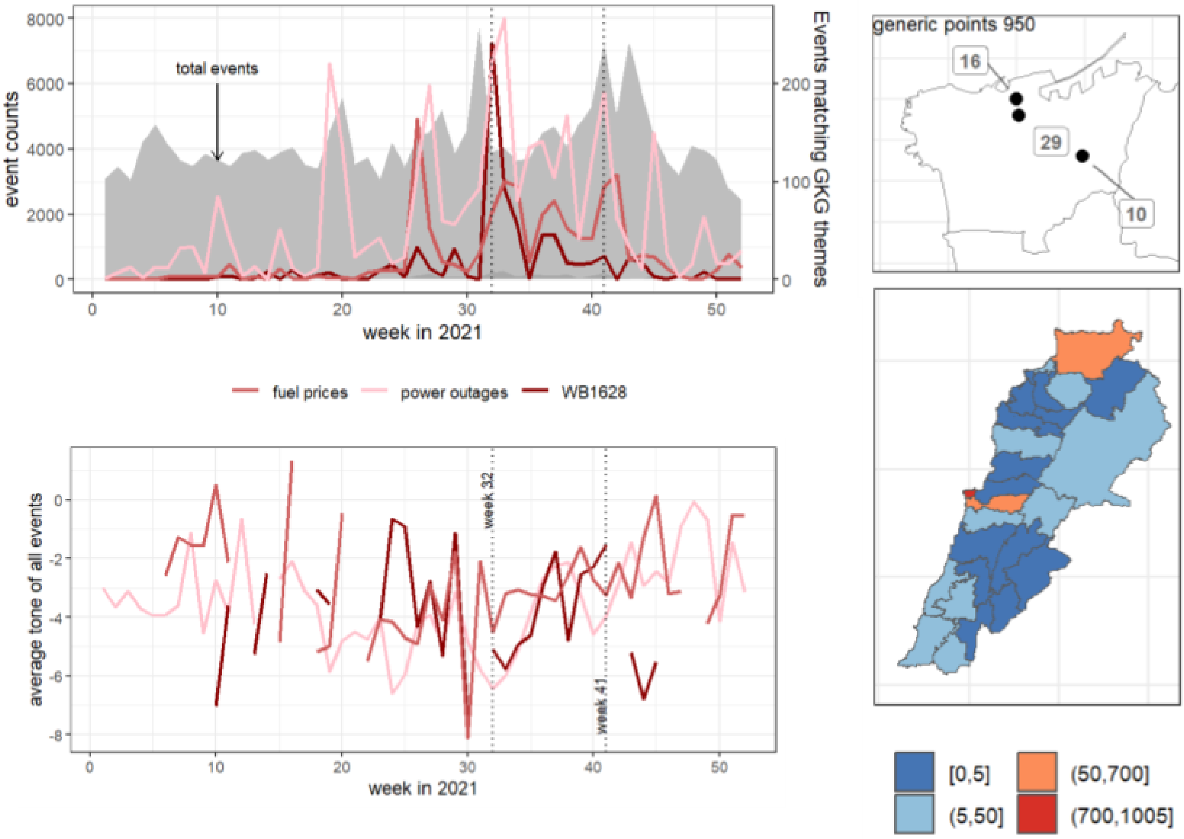
country (SBS News 2019). Roads were blocked (Ya Libnan 2019), and a general strike was declared (Osama 2019), crippling the nation's regular operations. While the majority of demonstrations remained peaceful, occasional clashes with security forces and counter-protesters emerged (The New Arab 2019). Amid this unprecedented unity and pressure, on 29 October, Prime Minister Saad Hariri announced his resignation (Al Arabiya News 2019; Dean 2019), marking a significant initial victory for the protest movement.

3.4 Case 4: Fuel Crisis 2021

The Lebanese fuel crisis of 2021 (FC hereafter) unfolded over several months, marked by a relentless series of severe fuel shortages that sent shockwaves throughout the nation. This crisis plunged the country already struggling with a confluence of crises (October Revolution, economic/banking crises, Covid-19, Port Beirut explosion, triple-digit inflation, and so forth) into a state of chaos in 2021, characterised by interminable queues at petrol stations and extended periods of darkness as power supply from private generators dwindled (Dagher, Jamali and Abi Younes 2023). While these events made headlines in the news, daily commodities became scarce and inaccessible to residents in Lebanon. Subsidies were being lifted for a variety of commodities including fuel, in summer 2021 (*ibid*.).

The situation painted a grim picture of a country grappling with multifaceted challenges, from economic turmoil to governance deficiencies. These events were covered extensively in the media including the international media, which means that GDELT is an appropriate source of information for mapping the events as they unfolded. However, FC was not an isolated crisis that the country was in the midst of during 2021. It follows that the identifying news reports that directly addressed FC, but not the other crisis, was challenging.

Figure 3.4 Events and GKG data related to the Fuel Crisis 2021



Source: Authors' own.

The approach we took to identify FC-relevant events was to focus on three relevant issues that could be directly and unambiguously connected to FC (the relevant GKG topics in brackets):

1. The suspension of the fuel subsidy (WB_1628_FUEL_SUBSIDIES)
2. Fuel prices (FUELPRICES)
3. Power outages (POWER_OUTAGE).

The time series graphs in Figure 3.4 provide a picture of the evolution of these events as captured in the media. Similar to the other case studies, the total event count (grey area graph) is not very helpful. While some of the peaks in the total events coincided with important moments in the FC story, others can be misleading. The peak in week 31 in total events, for example, had very little direct bearing on FC. Week 31 in 2021 is the first-year anniversary of the Beirut blast and the peak reflects the media coverage of it.

Figure 3.4 suggests that week 32 is a pivotal point in the fuel subsidy-related events in 2021.[20] This peak coincides with the central bank's announcement on 11 August 2021 that it would allow the import of fuel at the market price for the Lebanese pound. The importing at the pegged exchange rate was an effective fuel subsidy which the government had been reducing over the weeks/months preceding week 32. Smaller peaks in the WB1628 graph before week 32 coincide with these. But the change in week 32 was by far the worst in 2021, leading to a more than fourfold increase in fuel prices at the pump. Week 32 also marks the approximate start of a period where more news is coded with the GKG themes 'FUELPRICES' and 'POWER_OUTAGE' in the GDELT data set. For example, the peak in power outage reporting in October 2021 (week 41) coincides with the country-wide power blackout following the shutdown of the two largest power stations, the Zahrani and the Deir Ammar power stations. The maps in Figure 3.4 capture the geographic distribution of the WB1628 events and the fuel price-related events.

The aforementioned GDELT codes were used to filter news reports for manual checking. The filtered URLs for week 32 were for the most part related to FC. Most URLs in this case, if not all accessible ones, lead to articles about the fuel crisis, the lifting of fuel subsidies, increases in fuel prices, and protests that occurred in response to the fuel crisis at the time. Many inaccessible URLs (redirected to a landing page, generating errors, or just missing) also contained related terms as part of their URL which strongly suggested that at the time of their publication, those would also have been FC relevant.[21] In this case, GDELT data allows researchers to instantaneously access a wide range of articles published worldwide under the search topic, with appropriate filtering, and with respect to real-time events.

All three time series can also be used jointly in the FC narrative. For example, peaks in high power outages (in weeks 19 and 27) may indicate signs of the public utility, Electricité du Liban, struggling to procure fuel at subsidised prices. That power outages are a crucial factor that forced the government to abandon the fuel subsidies (week 29) is, to an extent, substantiated by the sequencing of the peaks in these event types. Also, abandoning the fuel subsidy leads to high fuel prices. The abandonment of the subsidy then led to a period of further fuel price-related events which are again FC-relevant. Therefore, this case study is a good example of how multiple GKG topics could be used to trace different

[20] In week 32, there were 241 events in new reports coded to 'WB_1628_FUEL_SUBSIDIES' in the GKG database.

[21] Keywords such as fuel, subsidy, outage, etc. were clearly relevant to FC. The following is an example of an inaccessible URL which is highly likely to have been FC-relevant at the time of the publication: **https://wacotrib.com/news/world/lebanon-increases-fuel-prices-by-more-than-35-amid-crisis/article_8108f692-fa8e-5d71-b527-b42737b054be.html**.

elements of the same story. Even though it is a messy data set, at least the quantitative analysis of it seems to make sense.

Next, we look at qualitative data. As for URLs from week 41, the GDELT codes used allow access to FC-relevant news articles for power outages, highlighting the nation during blackouts on that week (Gavlak 2021; PressTV 2021). The peak of week 41 (from 11 October to 17 October) highlights other issues that emerge from FC, such as its impact on schools, increased violence, and even people fleeing to other countries. Such offshoots of FC are unearthed during the manual checking of URLs. It is also possible to quantitatively check some of these tangential developments through GKG topics. However, we did not do so as it is beyond the scope of the present research. Overall, the data extracted using GDELT codes demonstrates the potential this tool has in yielding an extensive range of accurate results that bridge the gap between data and real-world events, all the while providing efficient access to researchers.

4. Discussion and conclusions

A unique feature of the methodology used in this research was the strong thread of academic–practitioner collaboration that runs through it. From the development of the grant application to the interpretation of results that are presented in this section, a range of key activities were guided by this collaboration.[22] This approach, we believe, will lean towards making this work relevant to both practitioners and academics alike. The interrogation of GDELT data for the four case studies in Lebanon was completed using this collaborative approach. The key findings of this work can be summarised as follows.

As explained at the outset, the use of GDELT for the purpose that we have described– as a more nuanced, disaggregated, case-specific data source – is new. The results suggest that, to a large extent, the data on **Event counts** and **Event tones** were helpful in charting key moments in known historical cases. The use of **Event data** critically depends on identifying a precise set of **GKG themes** to filter them with. **Events codes** can also be useful, especially if combined with appropriate **GKG themes**, as illustrated in Case 2 (RR).[23] Overall, however, the study revealed that the **GKG themes** are better suited to meet the research and practice needs of development studies. This is because the GKG topics were sufficiently nuanced to be able to map to common development topics. The topical taxonomy of the World Bank Group is a good example; three of our case studies used it for filtering events. And we agree with Czvetkó *et al*. (2021) that WB codes are a good choice for filtering events for development studies. In summary, a significant highlight of this research is that GDELT events filtered through the GKG themes accurately highlighted the important dates of our stories. This is a remarkable outcome given that some of our case studies unfolded amid an intricate mesh of unrelated activities.

That the filtered events helped identify relevant key moments in the cases does not mean that during this research we were able to confirm all those events as related/relevant to the case studies. We certainly tried to do this by reading the original news reports accessed through the URLs provided by GDELT. To a large degree, this was not successful because many of the URLs had already expired. In a limited number of cases, this meant that the expired URL was redirected to a new one where the original news report was archived. But in the majority of cases, the expired URLs were simply missing or redirected to a landing page of the new site with no trace of the original news item.

[22] In fact, at the institutional level, this collaboration extends to even earlier than the present project.

[23] In comparison to GKG themes, the Events codes which are based on the conflict-related CAMEO framework for coding was of limited relevance to four case studies in this research.

We used Google's **URL Fetch Service** to identify valid URLs. However, this approach did not fully serve the purpose as it could not identify the URLs that were redirected to a landing page. This meant that the researcher had to manually confirm such cases. The problem of missing URLs significantly affected our ability to confirm what events or news items were driving identified important dates. If there is a peak in events counts and if a large number of events from that peak were linked to a single news item (appearing in different news reports), that would help us postulate possible reasons for the observed peak. With so many expired ULRs, this was difficult to do. Therefore, we used a combination of three approaches to explain key dates/weeks in our case studies: (1) use secondary sources, (2) check the smaller set of working URLs, and (3) check if the full set of URLs (including the expired ones) contained keywords that were relevant to the case. This approach, while informative, was prohibitively labour-intensive. So further work is needed to streamline this process. Obviously, the issue of missing URLs is more pronounced for older data in GDELT. Conversely, real-time data or more recent data will have a negligible number of missing URLs.

The filtered event data was also interrogated through maps. A significant subset of events was further filtered from this analysis because the relevant news items did not reference a subnational geolocation. The maps and geolocalisation of events offered a quick snapshot of where the events were taking place across the country. An important caveat here is that many of the map-worthy events were geolocated to the capital as the seat of government/where statements are made and not necessarily where the event itself was taking place. Even after discounting for these data issues, the four cases that we investigated could be mapped in some detail. This would be of academic interest as a source of information to piece together the geographic spread of the case study as depicted in the media. Humanitarian/development practitioners interested in the geographic spread of topics, as captured in the media, would find this relevant.

The case studies illustrate that GDELT's datafied media reports may be collated and summarised to reveal trends in media coverage of specific topics and their tone; real-time if need be. The results indicate that the average tone of news items in particular can be a reasonably accurate barometer of where public opinion on specific topics is travelling. In some circumstances, this may constitute valuable information for action; the question is, in what circumstances? For decision makers and communication strategists, particularly in the development sector, this type of evidence can offer valuable insights into shaping media narratives and influencing public discourse. Lock (2020) used GDELT for a similar purpose, with a focus on global scale communications. However, she highlighted the potential for using GDELT for more local-scale communications. The results from the case studies seem to offer preliminary indications that this may be possible with GDELT data, at least at the country level. Such use of GDELT to support communication in research/practice could also be supported by its cross-domain capabilities. For example, by integrating data from multiple sectors and disciplines as we did in Case 4 (FC). This can enable researchers to explore the interconnectedness between different aspects of development, such as health, education, environment, and economics, and identify potential synergies and trade-offs in the context of the analysis of media discourses.

GDELT's potential in processing real-time event data raises the intriguing prospect of its application in early warning systems for development challenges, including but not limited to, natural disasters, conflicts, and economic crises. While this study did not identify a definitive approach to implement such a system, the avenue remains a compelling area for future exploration. The limited cross-domain work mentioned above may be an indication that common big data approaches in forecasting using the temporal association across different variables is possible. However, establishing such patterns could work better at scale rather than at the case-specific level used here. For example, when the filtered GDELT events data were rather small in number. This is because the particular GKG codes that we used were found in a limited number of cases in Lebanon. The strength in big data forecasting/prediction is the number of datapoints. So looking at the case studies, it would seem that an early warning system for development should instead be based on using GDELT data at a regional or global level.

Annexe A: Quantitative methodology

1. Data collection

The GDELT 2.0 Event Database and Global Knowledge Graph (GKG) are accessed in Google BigQuery. The '_partitioned' versions of events, eventmentions, and GKG tables are used to minimise the size of data processed by each search query. The tables are filtered to extract only the web articles[24] mentioning events that took place in Lebanon[25] within a specified period.[26] The main reason for filtering the eventmentions by MentionType 1 = web is that `MentionIdentifier` for other types are not distinct like a URL but are common strings as shown in Table A2. The count of non-web mentions is extremely low, and their exclusion is thus unlikely to cause any bias. Note that the EventCode column in the events table uses the CAMEO taxonomy, which is primarily focused on conflict-related events. This limitation necessitates the merging of the GKG table for filtering events to just those mentioned in articles that focus on themes from a broader range of categories.

2. Data integration

2.1 The filtered Events table is merged with the GKG table using the GDELT Events Mentions table as a connecting step.[27] This merging results in a data set containing one row for each article mentioning events that took place in Lebanon over the specified period. The results are saved as a BigQuery table to be used in the subsequent steps, thus minimising the compute costs of those queries. As a reference, the query step1_merged_events_mentions_gkg (see Annexe B) processes 1.9 TB each time it is run and costs US$11.875 (computed at the rate of US$6.25 per TB).

2.2 This data set is then grouped by aggregating[28] all mentions and GKG columns into single string values such that all mentions of an event are in a single row. The integration process results in a final data set containing events that occurred in Lebanon within the specified period,

[24] `MentionType = 1.`

[25] `ActionGeo_CountryCode  = 'LE'.`

[26]
```
_PARTITIONTIME >= TIMESTAMP("<PeriodStartDate>")
AND _PARTITIONTIME <= TIMESTAMP("<PeriodEndDate>").
```

[27] Right join filtered eventmentions to filtered events on `GLOBALEVENTID`, then right join filtered GKG on `MentionIdentifier` from mentions and `DocumentIdentifier` from GKG.

[28] All eventmentions and GKG table columns are aggregated over the group (GLOBALEVENTID) into string aggregation, separating each value by ';\n' characters. E.g.
`STRING_AGG(V2Themes,';\n') AS V2Themes.`

along with more detailed thematic coded categories from the GKG database. This integration allows for the inclusion of events published in articles that are coded with specific thematic categories, such as codes starting with 'WB_' (World Bank Group topical words) or CrisisLex taxonomy codes. Furthermore, this merging of the tables gives us access to all documents mentioning an event and filters all these document identifiers (i.e. URLs) based on keywords relevant to each case study.[29]

3. Weekly event counts

Generate weekly counts of events in the data set that cover the specific themes relevant to a case study. Count for each week the number of events that are mentioned in web articles related to the selected themes based on the coded categories from the GKG database: `V2Themes`. This step provides an overview of the volume of coverage on the selected themes over time. The event counts summarised at the weekly level mean that the resulting file has a small download size compared to step 2.

4. Identification of peaks

This uses the weekly counts of events that match certain GKG themes and keywords in mentioning document URLs. We use these to identify peaks, i.e. weeks with significantly higher event counts compared to the surrounding weeks. By zooming into the weeks with peaks, focus on the groups of documents that were featured during those periods. This allows for a more detailed examination of the events and articles that contributed to the peaks, providing insights into the specific case study.

5. URL verification

The list of URLs containing the relevant keywords and associated with relevant GKG themes within the weeks highlighted in the previous step are extracted. Use an App Script which uses Google's **URL Fetch Service** to issue HTTP/HTTPS requests and get responses for each URL considering redirected URLs as well, then filter out URLs that are not functional. This step is necessary because many URLs become inaccessible or stop working over time, ensuring that only accessible URLs are considered for further analysis.

Note: Google has set a **daily limit** on its URL Fetch Service. So this API has a daily quota of URLs it is permitted to read (20,000 URLs) which can be a binding constraint in a high-volume exercise such as this.

[29] E.g. `STRING_AGG(DocumentIdentifier,';\n') LIKE "%<keyword>%"`.

6. Analysis and interpretation

Conduct a qualitative analysis of the documents within the verified URLs to gain a deeper understanding of the events and themes covered in the selected case study. Analyse the content of the articles, identify key narratives, and extract relevant information related to the development studies context in Lebanon. Interpret the findings, draw conclusions, and discuss the implications of the selected events and themes within the broader context of development studies.

It is important to note that the exact details of the tools and scripts used for data integration and URL verification may vary depending on the specific resources available and the researchers' preferences.

Case Study 1: Garbage Crisis summer 2015

Related GKG V2Themes from World Bank Taxonomy

Table A1 Mention counts for each waste-related WB theme for the given time periods in 2015 in Lebanon

Theme	Week 31	Week 35	Week 38	2015 Total
WB_1797_SOLID_WASTE	346	363	396	3,062

Source: Authors' own.

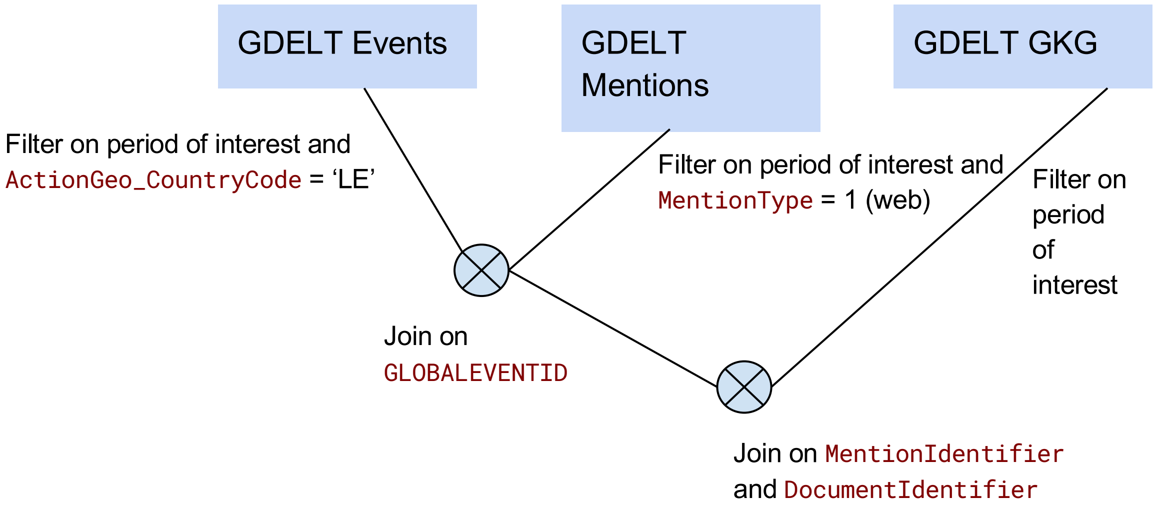
Table A2 Distribution of MentionType in the eventmentions_partitioned table for the period of 1 January 2015 to 31 December 2015

MentionType	Count	Example MentionIdentifier
1 (WEB)	10,243,61	http://www.timesofisrael.com/organizers-call-off-beirut-trash-protest-after-death/
2 (CITATIONONLY)	4,244	Al-Sharq al-Awsat website, London/BBC Monitoring/(c) BBC

Filters	events	eventmentions	GKG
_PARTITIONTIME >= TIMESTAMP("2015-01-01") AND _PARTITIONTIME <= TIMESTAMP("2015-12-31")	✓	✓	✓
ActionGeo_CountryCode = 'LE'	✓		
MentionType = 1		✓	

Source: Authors' own.

Figure A1 The merging of the three GDELT files


GDELT Events
GDELT Mentions
GDELT GKG
Filter on period of interest and ActionGeo_CountryCode = 'LE'
Filter on period of interest and MentionType = 1 (web)
Filter on period of interest
Join on GLOBALEVENTID
Join on MentionIdentifier and DocumentIdentifier

Source: Authors' own.

Annexe B: BigQuery SQL Code

B.1. step1_merged_events_mentions_gkg

```
SELECT E.*, EM.*, D.*
  FROM `gdelt-bq.gdeltv2.events_partitioned` E
  RIGHT JOIN `gdelt-bq.gdeltv2.eventmentions_partitioned` EM
  ON E.GLOBALEVENTID = EM.GLOBALEVENTID
  RIGHT JOIN `gdelt-bq.gdeltv2.gkg_partitioned` D
  ON EM.MentionIdentifier = D.DocumentIdentifier
  WHERE
    E._PARTITIONTIME >= TIMESTAMP("2015-01-01 00:00:00")
    AND E._PARTITIONTIME <= TIMESTAMP("2015-12-31 23:59:59")
    AND E.ActionGeo_CountryCode  = 'LE'
    AND EM._PARTITIONTIME >= TIMESTAMP("2015-01-01 00:00:00")
    AND EM._PARTITIONTIME <= TIMESTAMP("2015-12-31 23:59:59")
    AND EM.MentionType = 1 # WEB only
    AND D._PARTITIONTIME >= TIMESTAMP("2015-01-01 00:00:00")
    AND D._PARTITIONTIME <= TIMESTAMP("2015-12-31 23:59:59")
  ORDER BY E.GLOBALEVENTID, EM.MentionIdentifier, E.SQLDATE
```

B.2. step2_grouped_by_eventid

```
SELECT
  GLOBALEVENTID,
  EXTRACT(ISOWEEK FROM PARSE_DATE('%Y%m%d', Cast(MAX(SQLDATE) AS String))) AS
WeekNumber, #ISOWEEK
  STRING_AGG(V2Themes, ';\n') AS V2Themes,
  STRING_AGG(ActionGeo_FullName, ';\n') As ActionGeo_FullName,
  STRING_AGG(DocumentIdentifier,';\n') AS URL,
  COUNT(DISTINCT DocumentIdentifier) AS URLCount, #MentionsCountForEventID

  CASE WHEN REGEXP_CONTAINS(STRING_AGG(V2Themes, ';'),"WB_[0-9][0-9][0-
9]+_([a-zA-Z_]+)?WASTE([a-zA-Z_]+)?") THEN 1 ELSE 0 END AS ContainsWBTheme,
#theme contains pattern "WB_###_...WASTE..""
  CASE WHEN STRING_AGG(V2Themes, ';') LIKE "%PROTEST,%" THEN 1 ELSE 0 END AS
ContainsProtestTheme,
  #***Note-The theme is taken from GKG, and GKG is about a document. We can
only say that the events extracted correspond to a document that is related to
these themes:
  CASE WHEN STRING_AGG(DocumentIdentifier,';') LIKE "%garbage%" THEN 1 ELSE 0
END AS ContainsGarbage, #URL contains "garbage"
  CASE WHEN STRING_AGG(DocumentIdentifier,';') LIKE "%stink%" THEN 1 ELSE 0
END AS ContainsStink, #URL contains "stink"
  CASE WHEN STRING_AGG(DocumentIdentifier,';') LIKE "%waste%" THEN 1 ELSE 0
END AS ContainsWaste, #URL contains "waste"
  #Similarly, we can only say that the events extracted correspond to a
document that is related to these keywords:

  #events table columns - as these are common, taking MAX/MIN/.. works
  MAX( SQLDATE ) AS SQLDATE ,
  MAX( MonthYear ) AS MonthYear ,
  MAX( Year ) AS Year ,
  MAX( FractionDate ) AS FractionDate ,
  MAX( Actor1Code ) AS Actor1Code ,
  MAX( Actor1Name ) AS Actor1Name ,
  MAX( Actor1CountryCode ) AS Actor1CountryCode ,
  MAX( Actor1KnownGroupCode ) AS Actor1KnownGroupCode ,
  MAX( Actor1EthnicCode ) AS Actor1EthnicCode ,
  MAX( Actor1Religion1Code ) AS Actor1Religion1Code ,
  MAX( Actor1Religion2Code ) AS Actor1Religion2Code ,
  MAX( Actor1Type1Code ) AS Actor1Type1Code ,
  MAX( Actor1Type2Code ) AS Actor1Type2Code ,
  MAX( Actor1Type3Code ) AS Actor1Type3Code ,
  MAX( Actor2Code ) AS Actor2Code ,
  MAX( Actor2Name ) AS Actor2Name ,
  MAX( Actor2CountryCode ) AS Actor2CountryCode ,
  MAX( Actor2KnownGroupCode ) AS Actor2KnownGroupCode ,
  MAX( Actor2EthnicCode ) AS Actor2EthnicCode ,
  MAX( Actor2Religion1Code ) AS Actor2Religion1Code ,
  MAX( Actor2Religion2Code ) AS Actor2Religion2Code ,
  MAX( Actor2Type1Code ) AS Actor2Type1Code ,
  MAX( Actor2Type2Code ) AS Actor2Type2Code ,
  MAX( Actor2Type3Code ) AS Actor2Type3Code ,
  MAX( IsRootEvent ) AS IsRootEvent ,
  MAX( EventCode ) AS EventCode ,
  MAX( EventBaseCode ) AS EventBaseCode ,
  MAX( EventRootCode ) AS EventRootCode ,
  MAX( QuadClass ) AS QuadClass ,
  MAX( GoldsteinScale ) AS GoldsteinScale ,
  MAX( NumMentions ) AS NumMentions ,
  MAX( NumSources ) AS NumSources ,
  MAX( NumArticles ) AS NumArticles ,
  MAX( AvgTone ) AS AvgTone ,
  MAX( Actor1Geo_Type ) AS Actor1Geo_Type ,
  MAX( Actor1Geo_FullName ) AS Actor1Geo_FullName ,
  MAX( Actor1Geo_CountryCode ) AS Actor1Geo_CountryCode ,
  MAX( Actor1Geo_ADM1Code ) AS Actor1Geo_ADM1Code ,
  MAX( Actor1Geo_ADM2Code ) AS Actor1Geo_ADM2Code ,
  MAX( Actor1Geo_Lat ) AS Actor1Geo_Lat ,
  MAX( Actor1Geo_Long ) AS Actor1Geo_Long ,
  MAX( Actor1Geo_FeatureID ) AS Actor1Geo_FeatureID ,
  MAX( Actor2Geo_Type ) AS Actor2Geo_Type ,
  MAX( Actor2Geo_FullName ) AS Actor2Geo_FullName ,
  MAX( Actor2Geo_CountryCode ) AS Actor2Geo_CountryCode ,
  MAX( Actor2Geo_ADM1Code ) AS Actor2Geo_ADM1Code ,
  MAX( Actor2Geo_ADM2Code ) AS Actor2Geo_ADM2Code ,
  MAX( Actor2Geo_Lat ) AS Actor2Geo_Lat ,
  MAX( Actor2Geo_Long ) AS Actor2Geo_Long ,
  MAX( Actor2Geo_FeatureID ) AS Actor2Geo_FeatureID ,
  MAX( ActionGeo_Type ) AS ActionGeo_Type ,
  MAX( ActionGeo_FullName ) AS ActionGeo_FullName ,
  MAX( ActionGeo_CountryCode ) AS ActionGeo_CountryCode ,
  MAX( ActionGeo_ADM1Code ) AS ActionGeo_ADM1Code ,
  MAX( ActionGeo_ADM2Code ) AS ActionGeo_ADM2Code ,
  MAX( ActionGeo_Lat ) AS ActionGeo_Lat ,
  MAX( ActionGeo_Long ) AS ActionGeo_Long ,
  MAX( ActionGeo_FeatureID ) AS ActionGeo_FeatureID ,
  MAX( DATEADDED ) AS DATEADDED ,
  MAX( SOURCEURL ) AS SOURCEURL,

  #mentions table columns – aggregating values into a single string per column
  STRING_AGG(CAST(EventTimeDate as STRING),';\n') AS EventTimeDate,
  STRING_AGG(CAST(MentionTimeDate as STRING),';\n') AS MentionTimeDate,
  STRING_AGG(CAST(MentionType as STRING),';\n') AS MentionType,
  STRING_AGG(MentionSourceName,';\n') AS MentionSourceName,
  STRING_AGG(MentionIdentifier,';\n') AS MentionIdentifier,
  STRING_AGG(CAST(SentenceID as STRING),';\n') AS SentenceID,
  STRING_AGG(CAST(Actor1CharOffset as STRING),';\n') AS Actor1CharOffset,
  STRING_AGG(CAST(Actor2CharOffset as STRING),';\n') AS Actor2CharOffset,
  STRING_AGG(CAST(ActionCharOffset as STRING),';\n') AS ActionCharOffset,
  STRING_AGG(CAST(InRawText as STRING),';\n') AS InRawText,
  STRING_AGG(CAST(Confidence as STRING),';\n') AS Confidence,
  STRING_AGG(CAST(MentionDocLen as STRING),';\n') AS MentionDocLen,
  STRING_AGG(CAST(MentionDocTone as STRING),';\n') AS MentionDocTone,
  STRING_AGG(MentionDocTranslationInfo,';\n') AS MentionDocTranslationInfo,
  STRING_AGG(Extras,';\n') AS mentionsExtras,

  #gkg table columns - aggregating values into a single string per column
  STRING_AGG(GKGRECORDID,';\n') AS GKGRECORDID,
  STRING_AGG(CAST(DATE as STRING),';\n') AS DATE,
  STRING_AGG(CAST(SourceCollectionIdentifier as STRING),';\n') AS
SourceCollectionIdentifier,
  STRING_AGG(SourceCommonName,';\n') AS SourceCommonName,
  STRING_AGG(DocumentIdentifier,';\n') AS DocumentIdentifier,
  STRING_AGG(Counts,';\n') AS Counts,
  STRING_AGG(V2Counts,';\n') AS V2Counts,
  STRING_AGG(Themes,';\n') AS Themes,
  STRING_AGG(V2Themes,';\n') AS V2Themes,
  STRING_AGG(Locations,';\n') AS Locations,
  STRING_AGG(V2Locations,';\n') AS V2Locations,
  STRING_AGG(Persons,';\n') AS Persons,
  STRING_AGG(V2Persons,';\n') AS V2Persons,
  STRING_AGG(Organizations,';\n') AS Organizations,
  STRING_AGG(V2Organizations,';\n') AS V2Organizations,
  STRING_AGG(V2Tone,';\n') AS V2Tone,
  STRING_AGG(Dates,';\n') AS Dates,
  STRING_AGG(GCAM,';\n') AS GCAM,
  STRING_AGG(SharingImage,';\n') AS SharingImage,
  STRING_AGG(RelatedImages,';\n') AS RelatedImages,
  STRING_AGG(SocialImageEmbeds,';\n') AS SocialImageEmbeds,
  STRING_AGG(SocialVideoEmbeds,';\n') AS SocialVideoEmbeds,
  STRING_AGG(Quotations,';\n') AS Quotations,
  STRING_AGG(AllNames,';\n') AS AllNames,
  STRING_AGG(Amounts,';\n') AS Amounts,
  STRING_AGG(TranslationInfo,';\n') AS TranslationInfo,
  STRING_AGG(Extras,';\n') AS gkgExtras

FROM `bidds-375710.Lebanon_garbage2015.merged_events_mentions_gkg`

GROUP BY GLOBALEVENTID
```

B.3. step3A_weekly_counts

```
SELECT  Year, WeekNumber,

        #Event Counts
        SUM(ContainsWBTheme) AS CountContainsWBTheme,
        SUM(ContainsProtestTheme) AS CountContainsProtestTheme,
        SUM(CASE WHEN (ContainsGarbage + ContainsStink + ContainsWaste)>0 THEN
1 ELSE 0 END) AS CountContainsKeyWords,
        SUM(CASE WHEN ((ContainsGarbage + ContainsStink + ContainsWaste +
ContainsWBTheme)>0) THEN 1 ELSE 0 END) AS TotalKeywordsWBThemes,
        SUM(CASE WHEN ((ContainsGarbage + ContainsStink + ContainsWaste +
ContainsWBTheme + ContainsProtestTheme)>0) THEN 1 ELSE 0 END) AS
TotalAnyCriteria,

        #Sums of Tone - divide by relevant event count to get the average
        SUM(CASE WHEN ContainsWBTheme>0 THEN AvgTone ELSE 0 END) AS
ToneSumContainsWBTheme,
        SUM(CASE WHEN ContainsProtestTheme>0 THEN AvgTone ELSE 0 END) AS
ToneSumProtestTheme,
        SUM(CASE WHEN (ContainsGarbage + ContainsStink + ContainsWaste)>0 THEN
AvgTone ELSE 0 END) AS ToneSumContainsKeyWords,
        SUM(CASE WHEN ((ContainsGarbage + ContainsStink + ContainsWaste +
ContainsWBTheme)>0) THEN AvgTone ELSE 0 END) AS ToneSumKeywordsWBThemes,
        SUM(CASE WHEN ((ContainsGarbage + ContainsStink + ContainsWaste +
ContainsWBTheme + ContainsProtestTheme)>0) THEN AvgTone ELSE 0 END) AS
ToneSumAnyCriteria,

        COUNT(*) AS AllEventsCount,
        SUM(URLCount) AS TotalURLCount
```

```
FROM    `bidds-375710.Lebanon_garbage2015.grouped_by_eventid`

GROUP BY Year, WeekNumber
ORDER BY Year, WeekNumber
```

B.4. step3B_peak_week_check

```
SELECT Year, WeekNumber, GLOBALEVENTID, SQLDATE, EventCode, EventRootCode,
singleURL, URLCount,
ContainsWBTheme, ContainsProtestTheme, ContainsStink, ContainsWaste,
ContainsGarbage
FROM `bidds-375710.Lebanon_garbage2015.grouped_by_eventid`,
UNNEST(SPLIT(URL,";\n")) as singleURL
WHERE ContainsWBTheme>0
AND Year=2015 AND WeekNumber IN (31,35,38)
ORDER BY Year, WeekNumber,GLOBALEVENTID
```

```
SELECT
GLOB_QUERY.THEME AS THEME, GLOB_QUERY.WeekNumber AS WeekNumber, COUNT(*) AS
Mention_Count
FROM
(
SELECT WeekNumber,GLOBALEVENTID,
V2Themes,REGEXP_EXTRACT(themes,r'(^.[^,]+)') AS THEME
FROM `bidds-375710.Lebanon_JanDec2015.grouped_by_eventid`,
UNNEST(SPLIT(V2Themes,";")) AS Themes
WHERE ContainsWBTheme>0
)GLOB_QUERY
WHERE REGEXP_CONTAINS(GLOB_QUERY.THEME,"WB_[0-9][0-9][0-9]+_([a-zA-
Z_]+)?WASTE([a-zA-Z_]+)?")
GROUP BY GLOB_QUERY.THEME, GLOB_QUERY.WeekNumber
ORDER BY GLOB_QUERY.WeekNumber
```

Annexe C: Case study selection

These case studies show ways the GDELT data could be used to garner information about the way the cases were depicted in the news. The case studies involve a detailed mixed-method analysis of big data products related to humanitarian/development relevant case studies from Lebanon. The selection process and the final selection of cases were vetted by ACTED and World Vision (WV). This made sure that the selected case studies, and by extension any data/information about them, are pivotal to the work of humanitarian/development actors in Lebanon. There are other criteria in the list below, but this is clearly the most important.

The selection criteria were:

1. **What:** An **event or a trend** which is bounded by time. A time-bound event will help us easily separate GDELT information into three mutually exclusive segments: before, during, and after the event.
2. **Who:** The event or the trend involves **urban poor, refugees, or migrants** as the perpetrator or as an affected party.
3. **Why:** Because the event/trend has had an **impact/effect on humanitarian/development work** undertaken by practitioners working in Lebanon. As noted above, this is perhaps the most important selection criteria. Even though WV and ACTED are doing the selection, it is important that the event/trend has resonance with a wide cross-section of humanitarian/development (UN agencies, INGOs, NGOs, state, non-state, etc.) actors in Lebanon.
4. **When:** At present, the **GDELT 2.0** data streams only stretch back to 19 February 2015. Best if we can stick to this data set as it is the more comprehensive one. **GDELT 1.0** data stream goes back to 1979 but the data is quite sparse. If an event from before 19 February 2015 must be included, we can do so to illustrate the limits of GDELT 1.0.
5. **Where:** Local to **Lebanon**. Some cases may even be local to a city or a region in Lebanon. Let’s not consider international events/trends unless they have direct and unambiguous effects on local populations. This is primarily because we are filtering GDELT data at country level by looking at events that are geocoded for Lebanon.

References

Abumaria, D. (2019) ‘**Lebanese Blame Government for Blazes**’, [#879863320], *The Media Line*, 15 October (accessed 17 October 2023)

AFP (2018) ‘**Hundreds of Syrian Refugees Return Home from Lebanon**’, [#774026617], 23 July, *Digital Journal* (accessed 17 October 2023)

Al Arabiya News (2019) ‘**Lebanese PM Hariri Resigns After Weeks of Protests**’, [#881444083], *Al Arabiya English*, 29 October (accessed 17 October 2023)

Al Araby (2015) ‘**إغلاق “مطمر الناعمة” للنفايات في لبنان**’, [‘Closure of the Naameh Landfill in Lebanon’], [#450247639], 18 July (accessed 17 October 2023)

Al Bawaba (2015) ‘**Trash Near Beirut Airport Poses Safety Concerns for Air Travel**’, [#453141248], 29 July (accessed 17 October 2023)

Al Jazeera (2019) ‘**Protests Erupt in Lebanon Over Plans to Impose New Taxes**’, [#880448205], *Al Jazeera*, 17 October (accessed 17 October 2023)

Al-Manar (2015) ‘**جنة النفايات" الحكومية: لمزيد من العمل لحل الازمة- أرشيف موقع قناة المنار**’, [‘The Government “Waste Committee”: Further Work to Resolve the Crisis’], [#453333443], 29 July (accessed 17 October 2023)

Al-Manar TV Lebanon (2019) ‘**Lebanon Struggles to Battle Massive Wildfires**’, [#879771259], 15 October (accessed 17 October 2023)

Andrejevic, M. (2014) ‘**Big Data, Big Questions: The Big Data Divide**’, *International Journal of Communication* 8.17: 1673–89 (accessed 17 October 2023)

Baker, S.R.; Bloom, N. and Davis, S.J. (2016) ‘**Measuring Economic Policy Uncertainty**’, *The Quarterly Journal of Economics* 131.4: 1593–36 (accessed 17 October 2023)

Boyd, D. and Crawford, K. (2012) ‘**Critical Questions for Big Data Provocations for a Cultural, Technological, and Scholarly Phenomenon**’, *Information, Communication & Society* 15.5: 662–79 (accessed 17 October 2023)

Buckee, C.; Balsari, S. and Schroeder, A. (2022) ‘**Making Data for Good Better**’, *PLOS Digital Health* 1.1: e0000010 (accessed 17 October 2023)

Chehab, M. (2019) ‘**النائبة بولا يعقوبيان عن حرائق لبنان: البلد عاجز والدولة في غيبوبة**’, [‘Paula Yacoubian to Euronews About the Fires in Lebanon: The State is in a Coma and the Country is Helpless’], [879815117#], *Euronews*, 15 October (accessed 17 October 2023)

Chen, H.; Chiang, R. and Storey, V. (2012) ‘**Business Intelligence and Analytics: From Big Data to Big Impact**’, *MIS Quarterly* 36.4: 1165–88 (accessed 17 October 2023)

Cieslik, K. and Margócsy, D. (2022) ‘**Datafication, Power and Control in Development: A Historical Perspective on the Perils and Longevity of Data**’, *Progress in Development Studies* 22.4: 352–73 (accessed 17 October 2023)

Coniglio, N.D.; Peragine, V. and Vurchio, D. (2023) ‘The Effects of Refugees’ Camps on Hosting Areas: Social Conflicts and Economic Growth’, *World Development* 168: 106273

Couldry, N. and Mejias, U.A. (2020) *The Costs of Connection: How Data is Colonizing Human Life and Appropriating It for Capitalism*, Stanford CA: Stanford University Press

CSKC (2016) ***Social Movement Responding to the Lebanese Garbage Crisis***, Civil Society Knowledge Centre (accessed 17 October 2023)

Cukier, K. and Mayer-Schoenberger, V. (2013) ‘The Rise of Big Data: How It’s Changing the Way We Think About the World’, *Foreign Affairs* 92.3: 28–40

Czvetkó, T.; Honti, G.; Sebestyén, V. and Abonyi, J. (2021) ‘The Intertwining of World News with Sustainable Development Goals: An Effective Monitoring Tool’, *Heliyon* 7.2: e06174

Dagher, L.; Jamali, I. and Abi Younes, O. (2023) ‘**Extreme Energy Poverty: The Aftermath of Lebanon’s Economic Collapse**’, *Energy Policy* 183: 113783 (accessed 17 October 2023)

Dean, J.W. (2019) ‘**Breaking: Lebanese PM Hariri Announces Resignation Amid Large-Scale Protests**’, [#881330662], *Veterans Today*, 29 October (accessed 17 October 2023)

Diebold, F.X. (2021) ‘What’s the Big Idea? “Big Data” and Its Origins’, *Significance* 18.1: 36–37

Fakhoury, T. and Ozkul, D. (2019) ‘Syrian Refugees’ Return From Lebanon’, *Forced Migration Review* 62: 26–28

Francis, E. and Kanaan, A. (2019) ‘**Protests Sweep Lebanon as Fury at Ruling Elite Grows Over Economic Corruption**’, [#880715130], *Reuters*, 18 October (accessed 17 October)

Gamal-Gabriel, T. (2018) ‘**For Syrian Refugees, Fear of Conscription Prevents Return Home**’, [#767803330], *The Times of Israel*, 28 June (accessed 17 October)

Gavlak, D. (2021) ‘**Fire at Main Power Station, Blackouts Signal Lebanon’s Vulnerability**’, [#1008654673], *VOA*, 11 October (accessed 17 October)

GDELT (2021) ***New November 2021 GKG 2.0 Themes Lookup*** (accessed 26 October 2023)

Geha, C. and Talhouk, J. (2018) ***Politics and the Plight of Syrian Refugees in Lebanon***, Beirut: American University of Beirut (accessed 17 October)

GoL and United Nations (2018) ***Lebanon Crisis Response Plan 2017–2020*** (accessed 18 October)

Hatoum, B. (2018) ‘**Hundreds of Syrian Refugees Return Home From Lebanon**’, [#775365838], *Ottawa Citizen*, 28 July (accessed 18 October)

Hilbert, M. (2016) ‘Big Data for Development: A Review of Promises and Challenges’, *Development Policy Review* 34.1: 135–74

Honti, G.; Czvetkó, T.; Sebestyén, V. and Abonyi, J. (2021) ‘Data Describing the Relationship between World News and Sustainable Development Goals’, *Data in Brief* 36: 106978

HRW (2019) ‘Lebanon’, in ***World Report 2019***, New York NY: Human Rights Watch

Khatib, D.K. (2022) ‘17 October (2019) Revolution in Lebanon: A Preliminary Analysis’, in L. Issaev and A. Korotayev (eds), *New Wave of Revolutions in the MENA Region: A Comparative Perspective*, Cham: Springer

Khneisser, M. (2020) ‘The Specter of “Politics” and Ghosts of “Alternatives” Past: Lebanese “Civil Society” and the Antinomies of Contemporary Politics’, *Critical Sociology* 46.3: 359–77

Khneisser, M. (2019) ‘**The Marketing of Protest and Antinomies of Collective Organization in Lebanon**’, *Critical Sociology* 45.7–8: 1111–32 (accessed 18 October)

Kitchin, R. (2022) *The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures*, 2nd ed., Los Angeles CA: SAGE Publications

Kondraganti, A.; Narayanamurthy, G. and Sharifi, H. (2022) ‘**A Systematic Literature Review on the Use of Big Data Analytics in Humanitarian and Disaster Operations**’, *Annals of Operations Research* (accessed 18 October)

Kshetri, N.; Fredriksson, T. and Torres, D.C.R. (2017) *Big Data and Cloud Computing for Development: Lessons from Key Industries and Economies in the Global South*, New York NY and Abingdon: Taylor & Francis

Lakshman, R.W.D.; te Lintelo, D.J.H.; Mansour, W. and Ford, H. (2020) ***Making Use of Big Data to Inform Policy: Lebanon Case Study***, IDS Opinion, blog, 7 September (accessed 18 October)

LaValle, S.; Lesser, E.; Shockley, R.; Hopkins, M.S. and Kruschwitz, N. (2011) ‘Big Data, Analytics and the Path from Insights to Value’, *MIT Sloan Management Review* 52.2: 21–32

Levin, N.; Ali, S. and 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 (accessed 18 October)

Lock, I. (2020) ‘**Debating Glyphosate: A Macro Perspective on the Role of Strategic Communication in Forming and Monitoring A Global Issue Arena Using Inductive Topic Modelling**’, *International Journal of Strategic Communication* 14.4: 223–45 (accessed 18 October)

Makdisi, S. and Marktanner, M. (2009) ‘Trapped by Consociationalism: The Case of Lebanon’, *Topics in Middle Eastern and North African Economies* 11

Mann, L. (2018) ‘Left to Other Peoples’ Devices? A Political Economy Perspective on the Big Data Revolution in Development’, *Development and Change* 49.1: 3–36

Manyika, J. *et al*. (2011) ***Big Data: The Next Frontier for Innovation, Competition, and Productivity***, New York NY: McKinsey Global Institute (accessed 18 October)

Mayer-Schönberger, V. and Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Boston MA: Houghton Mifflin Harcourt

Mejias, U.A. and Couldry, N. (2019) 'Datafication', *Internet Policy Review* 8.4

Milan, S. and Treré, E. (2019) 'Big Data from the South(s): Beyond Data Universalism', *Television & New Media* 20.4: 319–35

Mroue, B. (2018) '**Hundreds of Syrians Leave Lebanon for Long-Awaited Reunions**', [#768391242], *The Times of Israel*, 30 June (accessed 18 October)

Naharnet Newsdesk (2018) '**Syrian Refugees Living in Arsal Start Returning to Syria**', *Naharnet*, 28 June (accessed 18 October)

Naharnet Newsdesk (2015a) '**Hbaline Residents to Block Roads Saturday as Minister Says Sukleen to Keep Collecting Waste in Beirut**', [#449787929], *Naharnet*, 16 July (accessed 18 October)

Naharnet Newsdesk (2015b) '**Residents of Jbeil and Hbaline Block Roads to Landfill**', [#450223056], *Naharnet*, 18 July (accessed 18 October)

Obeid, G. (2018) '**Lebanon Welcomes Russian Proposal on Refugees**', [#775322965], *Arab America*, 28 July (accessed 18 October)

Osama, S. (2019) '**Nationwide Strike in Lebanon on Fifth Day of Protests**', [#881212958], *Ahram Online*, 21 October (accessed 18 October)

PressTV (2021) '**Jordan Pledges Support for Crisis-Hit Lebanon**', [#1008654673], 11 October (accessed 18 October)

Qiao, F. *et al*. (2017) 'Predicting Social Unrest Events with Hidden Markov Models Using GDELT', *Discrete Dynamics in Nature and Society* 2017: 1–13

Rose, S. (2018) '**Syrian Refugees in Lebanon Alarmed by Russian Involvement in Planned Repatriation**', [#775411070], *The National*, 29 July (accessed 18 October)

Saad, A. (2018) '**ليوم.. 400 لاجئ سوري يعودون من لبنان إلى سوريا**', ['Today 400 Syrian Refugees Return from Lebanon to Syria'], [#767804461], *Al Bawaba*, 28 June (accessed 18 October)

SBS News (2019) '**Huge Crowds Chant for "Revolution" in Lebanon's Biggest Day of Protest Yet**', [#881057609], 20 October (accessed 18 October)

Schrodt, P.A. (2012) ***CAMEO Conflict and Mediation Event Observations Event and Actor Codebook*** (accessed 18 October)

Sharma, P. and Joshi, A. (2020) 'Challenges of Using Big Data for Humanitarian Relief: Lessons From the Literature', *Journal of Humanitarian Logistics and Supply Chain Management* 10.4: 423–46

Sumner, A. (2022) *What is Development Studies?*, Bonn: European Association of Development Research and Training Institutes

te Lintelo, D.J.H. *et al*. (2018) ***Wellbeing and Protracted Urban Displacement: Refugees and Hosts in Jordan and Lebanon***, Brighton: Institute of Development Studies (accessed 18 October)

The GDELT Project (2015) '**World Bank Group Topical Taxonomy Now in GKG**', 2 March (accessed 17 October)

The New Arab (2019) 'Security Forces Crack Down as Lebanon Protests Rage Against Ruling Elite', [#880565604], 19 October

Transparency International (2019) ***Corruption Perceptions Index 2018***, 29 January (accessed 18 October)

Turner, L. (2015) 'Explaining the (Non-)Encampment of Syrian Refugees: Security, Class and the Labour Market in Lebanon and Jordan', *Mediterranean Politics* 20.3: 386–404

UNHCR (2019) ***Lebanon: December 2018*** (accessed 18 October)

Williams, S. (2020) ***Exploration of the Global Database of Events, Language and Tone (GDELT), With Specific Application to Disaster Reporting***, Newport: Office for National Statistics (accessed 18 October)

World Bank (2020) ***Lebanon Economic Monitor, Fall 2019: So When Gravity Beckons, the Poor Don't Fall*** (accessed 18 October)

World Bank (2016) ***Theme Taxonomy and Definitions*** (accessed 18 October)

Yahya, M. (2020) '**All Fall Down**', *Carnegie Middle East Center*, 23 July (accessed 18 October)

Yahya, M. (2019) '**Out With the Old, In With What?**', *Carnegie Middle East Center*, 16 December (accessed 18 October)

Ya Libnan (2019) '**All Major Roads Leading to Beirut Airport Closed by Protesters**', [#880488849], 18 October (accessed 18 October)

Yonamine, J.E. (2013) 'A Nuanced Study of Political Conflict Using the Global Datasets of Events Location and Tone (GDELT) Dataset', PhD dissertation, Pennsylvania State University

institute of development studies

Delivering world-class research, learning and teaching that transforms the knowledge, action and leadership needed for more equitable and sustainable development globally.

Institute of Development Studies
Library Road
Brighton, BN1 9RE
United Kingdom
+44 (0)1273 606261
ids.ac.uk

Charity Registration Number 306371
Charitable Company Number 877338