

Xenophobic Events vs. Refugee Population – Using GDELT to Identify Countries with Disproportionate Coverage

Himarsha R. Jayanetti¹[0000–1111–2222–3333], Erika
Frydenlund²[0000–0002–7694–7845], and Michele C. Weigle¹[0000–0002–2787–7166]

¹ Old Dominion University, Norfolk, Virginia, USA
{hjaya002}@odu.edu {mweigle}@cs.odu.edu

² Virginia Modeling, Analysis and Simulation Center, Old Dominion University,
Suffolk, Virginia, USA
{efrydenl}@odu.edu

Abstract. In this preliminary study, we used the Global Database of Events, Language, and Tone (GDELT) database to examine xenophobic events reported in the media during 2022. We collected a dataset of 2,778 unique events and created a choropleth map illustrating the frequency of events scaled by the refugee population’s proportion in each host country. We identified the top 10 countries with the highest scaled event frequencies among those with more than 50,000 refugees. Contrary to the belief that hosting a significant number of forced migrants results in higher xenophobic incidents, our findings indicate a potential connection to political factors. We also categorized the 20 root event codes in the CAMEO event data as either “Direct” or “Indirect”. Almost 90% of the events related to refugees in 2022 were classified as “Indirect”.

Keywords: Xenophobia · Refugees and Migrants · GDELT · Big Data

1 Introduction

People move around the world in pursuit of better opportunities or to flee conflicts and natural disasters. There are 281 million international migrants, or one in every 30 people worldwide [9], and more than 82 million of them have been forcibly displaced [17]. These migrants make an effort to coexist in their host communities. However, widespread xenophobic and racist violence makes it difficult to uphold societal order and provide equal access to opportunities, resources, and even human dignity. Hence, it is imperative to study such xenophobic incidents and examine the underlying factors contributing to hostile behavior towards refugees in order to fight xenophobia.

In our preliminary study, we use a massive and regularly updated dataset of online, TV, and news reporting from Global Data on Events, Location, and Tone (GDELT)³ to look for ways to characterize xenophobic events. By collecting

³ <https://www.gdeltproject.org/>

data from 2022, we were able to analyze and derive meaningful insights from the dataset within that one-year period.

Despite GDELT’s richness as a data source, harnessing it for very specific tasks like tracking xenophobic events presents challenges. Hoffmann et al. [8] pointed out the challenges of utilizing big data news sources, such as the need for a deep understanding of the database’s complexity and a significant time investment in data analysis. Hammond and Weidmann [7] evaluated the applicability of GDELT data studying political violence. The results emphasize the need for caution when using GDELT for geospatial studies through a comparison with two hand-coded conflict event datasets. Hence, conducting exploratory research using a short period, such as a single year (2022), is crucial for better comprehension of the challenges involved.

In this paper, we are addressing two key research questions (RQ1 and RQ2):

1. RQ1: Can we identify countries with an unexpectedly high number of refugee-related events reported in the news media relative to their refugee populations?
2. RQ2: Can the 20 root event codes be broadly categorized into “Direct” and “Indirect”, and how prevalent are “Indirect” actions compared to “Direct” actions within our dataset?

Our findings challenge the notion that hosting a large number of forced migrants directly correlates with an increase in xenophobic incidents, suggesting a complex interplay of factors beyond refugee populations and highlighting the potential influence of political dynamics. Moreover, we were able to group the 20 root event codes into “Direct” and “Indirect” categories and found that a significantly higher proportion were classified as “Indirect” actions.

2 Background and Related Work

GDELT is a prime example of big data, with billions of records spanning from 1979 to the present. It uses the Conflict and Mediation Event Observations (CAMEO) taxonomy, a framework for coding event-related data, to automatically code data for use in research [6]. Researchers have used GDELT data for a variety of studies, such as studying the effects of civil unrest, complexity in terms of political activities, and capturing peace through the Global Peace Index (GPI) [1,19,22]. Vargo et al. studied the power of fake news from 2014 to 2016 in online news media using GDELT [18]. Their research revealed that although the prevalence of fake news has risen, these websites do not possess undue influence. Other researchers have used social media platforms as well as newspaper articles in opinion mining about a range of topics from online education to industrial production [5,16].

Various studies have examined forced migration and policy implications in countries that support migration [2,3,4]. Napierala et al. summarize how big data, like GDELT, have been used to help predict refugee flows and humanitarian situations [10]. Yesilbas et al. utilized GDELT to build a large dataset of global

news to study the tone, volume, and topics of media coverage of refugees [21]. They found that a negative tone occurred both because of the anti-migrant sentiment as well as sorrow and empathy for the refugees. Weisser used GDELT to characterize the actors (e.g., nonprofits, businesses, and government) that drive the conversation and support for refugees and migrants [20].

Our research shares similarities with previous studies in that we aim to analyze events related to refugee and migrant communities. Our approach differs from previous analyses and datasets in that our final objective is to develop a real-time monitoring system for xenophobic events utilizing GDELT data to identify potential hotspots of violence to predict potential escalation. While combating xenophobia is within the mandates of international organizations such as the United Nations High Commissioner for Refugees (UNHCR) and the International Organization for Migration (IOM), there is no worldwide tool for tracking these events. The Internal Displacement Monitoring Centre⁴ has designed a widely-adopted hand-coded data synthesis tool to monitor internal displacement caused by conflict and natural disasters. ACLED⁵ similarly tracks protest and violence across the world. Xenowatch⁶ is an online heatmap using data that researchers have hand-coded from user-submitted news articles of xenophobic events in South Africa. These tools have had immeasurable impacts on research and decision-making, but no such tool exists on the global scale for xenophobic events. The ultimate aim of this study is to construct such a tool; however, this paper discusses our preliminary exploratory analysis of 2022 global news data automatically coded by GDELT for events involving migrants.

3 Methodology

3.1 Understanding the GDELT Database

In our study, we are using the GDELT 2.0 database, which is updated every 15 minutes and translates articles from around the world from 65 different languages into English [12]. In our exploration of the use of GDELT, we identified three key tables in the database:

1. **Event:** Contains data about events happening globally. Each row represents a single event, coded with information such as an event identification number (`GLOBALEVENTID`), actors (`Actor1Code`, `Actor2Code`), and location (`Actor1CountryCode`, `Actor2CountryCode`).
2. **Event Mentions:** Contains a row for each mention of the event in a news article or other source. Each mention is coded with its respective `GLOBALEVENTID` and information about the tone of the mention (positive or negative). The `GLOBALEVENTID` corresponds to the `GLOBALEVENTID` from the Event Table, and the external identifier `MentionIdentifier` for the source document uniquely identifies the document.

⁴ <https://www.internal-displacement.org>

⁵ <https://acleddata.com/>

⁶ <https://www.xenowatch.ac.za>

3. **Global Knowledge Graph (GKG)** [15]: Connects data from various sources to form an interconnected network that encapsulates events around the world, their corresponding contexts, associated actors, and the overall sentiment of media coverage. The `DocumentIdentifier` field corresponds to the `MentionIdentifier` in the Event Mentions table. This GKG table comprises the results obtained from the Global Content Analysis Measures (GCAM) system, which employs multiple content analysis tools to capture over 2,230 latent dimensions for each news article monitored by GDELT [14].

3.2 Data Collection Methods and Criteria

We used the year 2022 for our analysis after considering several factors. Firstly, we aimed to focus on a recent timeframe with the intention of building upon our understanding and awareness of recent events. We intentionally avoided selecting the year 2021 due to COVID-19 disruptions, including the peak of the pandemic and related restrictions. We avoided recent years due to challenges in accounting for changes in migration patterns during the pandemic due to lockdown measures. By focusing on 2022, we achieved a balance between recentness and mitigating potential confounding factors that could affect our findings.

To capture xenophobic events in the media, we identified 8 main GKG themes [13] that are related to negative actions against immigrants, which we refer to collectively as “GKG_REF”:

- DISCRIMINATION_IMMIGRATION_XENOPHOBIA
- DISCRIMINATION_IMMIGRATION_ANTIIMMIGRANTS
- DISCRIMINATION_IMMIGRATION_OPPPOSED_TO_IMMIGRANTS
- DISCRIMINATION_IMMIGRATION_AGAINST_IMMIGRANTS
- DISCRIMINATION_IMMIGRATION_ATTACKS_ON_IMMIGRANTS
- DISCRIMINATION_IMMIGRATION_ATTACKS_AGAINST_IMMIGRANTS
- DISCRIMINATION_IMMIGRATION_XENOPHOBE
- DISCRIMINATION_IMMIGRATION_XENOPHOBES

Figure 1 depicts the steps of our data collection and filtering process. We used Google BigQuery⁷ to connect to the GDELT databases and collected data for 2022 having “GKG_REF” themes (using the query filter `V2Themes like DISCRIMINATION_IMMIGRATION`). This initial dataset comprised 601,855 records from the combined Event, Event Mentions, and GKG tables. After filtering for `Actor1Code` or `Actor2Code` corresponding to REF (actor code for "Refugee"), we obtained 9,392 records. We further filtered out entries without a country code, resulting in 5,448 records. Among these, we identified 2,778 unique events, which were saved as a CSV file with the country code and event frequency.

Our goal is to highlight those countries that have a greater number of refugee-related events reported in news media than would be expected based on the relative size of their refugee populations. This might point to countries where

⁷ <https://cloud.google.com/bigquery>

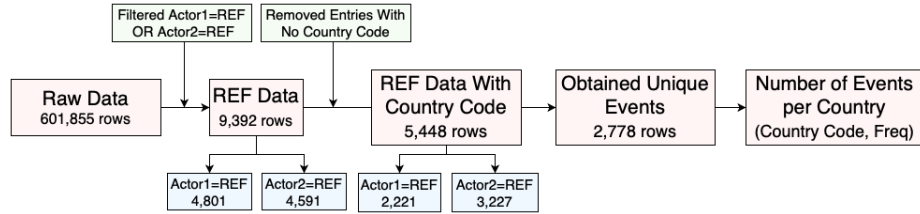


Fig. 1: Data analysis and filtering process

refugees have become a contentious political issue. To do this, we first calculated the proportion of refugees per host country by dividing the number of refugees by the country’s total population (RT). We obtained the population of displaced people (refugees, asylum seekers, and others in refugee-like situations) for each country of asylum (as of mid-2022) from UNHCR⁸ and the total country population for 2022 from the US Census Bureau.⁹ Then, we divided the number of relevant GDELT events per country by that country’s proportion of refugees. For instance, in 2022, Turkey and Germany had similar total populations (~ 80 million) and a similar number of relevant GDELT events (~ 100). But, since Germany’s proportion of refugees to the total population ($\sim 3\%$) was lower than Turkey’s ($\sim 5\%$), Germany was given a higher scaled frequency score.

4 Results

4.1 Scaled Frequency of Events Per Country

Figure 2 is a choropleth map showing the scaled frequency of the number of unique events in the data. Gray areas indicate countries without available data. Dark purple represents countries with the highest scaled event frequencies, meaning they had more relevant events than expected based on their proportion of refugees. Light blue represents countries with the lowest scaled event frequencies.

Of note in Figure 2 is that three countries – the USA, China, and Ukraine – produce the largest frequency of reported events of discrimination/violence against migrants relative to the volume of refugees hosted in their countries. This may be explained by one of the biases in news coverage of GDELT, where a significant number of media sources originate from the US and China. Additionally, Ukraine is the third largest and also most recent refugee situation in the world. The recency of this event may explain its prevalence in news coverage.

Table 1 shows the top 10 countries based on the scaled frequency of unique events. When reporting the top 10 countries, we established a threshold of hosting a minimum of 50,000 refugees and other displaced people. This decision was made to prevent a significant decrease in the RT value, which in turn would greatly increase the scaled frequency.

⁸ <https://www.unhcr.org/refugee-statistics/download/?url=c11N2c>

⁹ <https://www.census.gov/programs-surveys/international-programs/data.html>

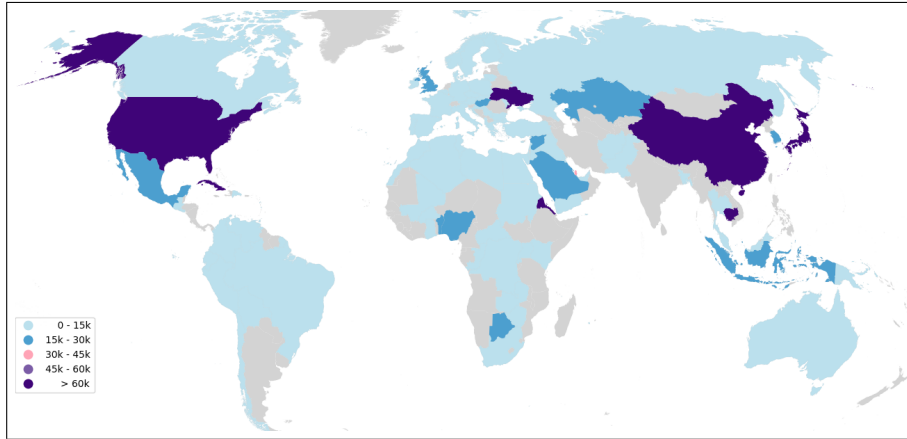


Fig. 2: Choropleth map illustrating the scaled frequency of unique events

According to the UNHCR, 36% of all displaced people are hosted in just five countries: Turkey, Colombia, Germany, Pakistan, and Uganda. Notably, none of these are in the top 10 by scaled frequency. Counter to the idea that hosting a large number of forced migrants increases xenophobic events, Nigeria (NGA) is in the top three countries, but the number of displaced people hosted is quite low. Rwanda (RWA) additionally has a disproportionate number of xenophobic events coverage given how small the national population is compared to others in the top 10 list. Together, these entries suggest that xenophobic events coverage is related to more than just the burden of hosting large numbers of refugees, but indicates a potential association with political dynamics.

Table 1: Top 10 countries based on the scaled frequency of unique events

Country Code (CC)	Event Frequency (F)	Refugee Population (RP)	Total Population (TP)	Refugee/Total Population (RT)	Scaled Frequency (F/RT)
USA	354	1,787,504	337,341,954	0.0053	66,807.71
GBR	158	359,311	67,791,400	0.0053	29,809.95
NGA	9	84,302	225,082,083	0.0004	24,029.55
MEX	59	498,213	129,150,971	0.0039	15,294.48
ITA	62	354,414	61,095,551	0.0058	10,687.85
CAN	28	126,499	38,232,593	0.0033	8,462.62
RUS	83	1,463,050	142,021,981	0.0103	8,057.02
ZAF	32	240,077	57,516,665	0.0042	7,666.43
RWA	73	128,056	13,173,730	0.0097	7,509.86
BRA	17	562,577	217,240,060	0.0026	6,564.58

4.2 Event Root Code Classification

To explore this further, we characterized the CAMEO event codes [11], which broadly label the news articles using action words describing the news coverage. These codes are organized into 20 root event codes, numbered from 01 to 20. We categorized these into two broad categories, “Direct” and “Indirect” actions. By indirect, we look at actions related to diplomacy and “talking” about an issue, and direct actions are those that require physical actions, such as “demanding” or even “fighting” or “assaulting”. Table 2 shows the classification of the 20 root codes into our two predefined categories. Out of the total 2,778 events, we classified 360 (13%) as “Direct” and 2,418 (87%) as “Indirect”.

Table 2: Cameo event root codes, its description, and our categorization

Indirect		Direct	
Event RootCode	Description	Event RootCode	Description
01	Make Public Statement	09	Investigate
02	Appeal	10	Demand
03	Express Intent to Cooperate	13	Threaten
04	Consult	14	Protest
05	Engage in Diplomatic Cooperation	15	Exhibit Force Posture
06	Engage in Material Cooperation	16	Reduce Relations
07	Provide Aid	17	Coerce
08	Yield	18	Assault
11	Disapprove	19	Fight
12	Reject	20	Use Unconventional Mass

Table 3 presents a breakdown of data by country code of the top 10 countries, specifically showcasing the number of events for the “Direct” and “Indirect” categories, as well as the total frequency of events. It captures the idea that much of the media coverage of discrimination and/or violence against refugees and other displaced people is largely “all talk and no action”. Assuming that an actual xenophobic action would fall under a “direct” event (which contains assault, fight, protest, and threaten actions), 2022 saw very few events in locations that have a large relative frequency of event reporting. Nigeria, for instance, is in the top 10 countries reporting the most on discrimination and/or violence against refugees but had zero “direct” events.

Further, potentially qualitative deep-analysis of newspaper articles is required to understand whether indirect actions are a predictor of future xenophobic events. What is clear is that actual events are disproportionately outnumbered by news coverage that talks about doing something. While large datasets like GDELT can provide evidence of these trends, further review of samples of the actual news articles is required to understand how these CAMEO codes may be refined and tested as predictors of upticks in direct xenophobic events.

Table 3: Number of “Indirect” and “Direct” events for the top 10 countries

Country Code	Indirect	Direct	Total
USA	298 (84%)	56 (16%)	354
GBR	142 (90%)	16 (10%)	158
NGA	9 (100%)	0 (0%)	9
MEX	52 (88%)	7 (12%)	59
ITA	49 (79%)	13 (21%)	62
CAN	25 (89%)	3 (11%)	28
RUS	73 (88%)	10 (12%)	83
ZAF	31 (97%)	1 (3%)	32
RWA	60 (82%)	13 (18%)	73
BRA	9 (53%)	8 (47%)	17

5 Future Work

When considering future work, we need to address some limitations in our study. Firstly, we only analyzed events with location information in the dataset. Therefore, it would be valuable to investigate events without location information and explore their characteristics and implications. Additionally, our study focused on a single year, specifically 2022. To gain a better understanding of patterns and phenomena, it is important to extend the study to a wider time range and assess the robustness and consistency of the identified findings. Furthermore, we can use the results of our study to develop improved interactive visualizations for monitoring xenophobic violence against refugees and migrants.

6 Conclusion

In this study, we used the GDELT database to explore and characterize xenophobic events reported in the media in 2022. Our goal was to highlight those countries that had a greater number of refugee-related events reported in news media than would be expected based on the relative size of their refugee populations (RQ1). We identified 8 GKG themes, collectively referred to as “GKG_REF” present in the GDELT dataset. Our dataset is comprised of 2,778 unique events involving “REF” actors (actor code for refugees in GDELT) having “GKG_REF” themes. The results of our analysis revealed countries with higher scaled event frequencies, indicating a greater number of xenophobic events reported in the media relative to their refugee populations. Notably, the United States, China, and Ukraine emerged as countries with a significant frequency of reported events despite variations in their refugee populations. We reported the top 10 countries with the highest scaled event frequencies (of those having more than 50,000 refugees). Counter to the idea that hosting a large number of forced migrants leads to an increase in xenophobic incidents, our findings showed otherwise. Our results suggested that xenophobic events coverage is related to more than just the burden of hosting large numbers of refugees, but indicating a potential association with political dynamics. Further, we grouped the 20 root event codes in

CAMEO event categories into two primary categories: “Direct” and “Indirect” actions (RQ2). Out of the total 2,778 events, 13% (360) were classified as “Direct”, while 87% (2,418) were categorized as “Indirect”, suggesting a greater prevalence of indirect actions like diplomacy and discourse compared to direct actions such as demanding or assaulting. If we consider that xenophobic actions are typically categorized as “direct” events, which include actions like assault, fight, protest, and threaten, there were only a few occurrences in 2022 in the top 10 countries compared to “indirect” events. By discussing the challenges associated with using extensive datasets like GDELT for specific tasks, we emphasized the significance of conducting focused exploratory research within a specific timeframe (such as a single year – 2022), to gain a deeper understanding of these challenges.

Acknowledgements This research was funded under the project “Data Science for Social Good: Mining and Visualizing Worldwide News to Monitor Xenophobic Violence”, through the 2022-2023 ODU Data Science Seed Funding Program.

References

1. Fang, P., Gao, J., Fan, F., Yang, L.: Identifying political “hot” spots through massive media data analysis. In: Proceedings of the 9th International Conference on Social, Cultural, and Behavioral Modeling (SBP-BRiMS). pp. 282–290 (2016). https://doi.org/10.1007/978-3-319-39931-7_27
2. Frydenlund, E., Jones, E.C., Padilla, J.J.: Mobility in crisis: An agent-based model of refugees’ flight to safety. *Human Simulation: Perspectives, Insights, and Applications* pp. 191–208 (2019). https://doi.org/10.1007/978-3-030-17090-5_11
3. Frydenlund, E., Padilla, J.J.: Opportunities and challenges of using computer-based simulation in migration and displacement research: A focus on Lesbos, Greece. In: Bradley, M., Milner, J. (eds.) *Documenting Displacement: Questioning Methodological Boundaries in Forced Migration Research*, chap. 11, pp. 279–308. McGill-Queen’s University Press (2022)
4. Frydenlund, E., Yilmaz Şener, M., Gore, R., Boshuijzen-van Burken, C., Bozdog, E., De Kock, C.: Characterizing the mobile phone use patterns of refugee-hosting provinces in Turkey. In: Salah, A., Pentland, A., Lepri, B., Letouzé, E. (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, pp. 417–431 (2019). https://doi.org/10.1007/978-3-030-12554-7_21
5. Fu, Z., Yan, M., Meng, C., Wang, W., Hu, X., Li, Y., Wang, J., He, Z., Wang, Z.: Opinion mining about online education basing on GDELT and Twitter data. In: Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). pp. 741–745 (2020). <https://doi.org/10.1109/WIIAT50758.2020.00114>
6. Gerner, D.J., Schrodtt, P.A., Yilmaz, O., Abu-Jabr, R.: Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. In: Proceedings of the Annual Meeting of the International Studies Association (2002)
7. Hammond, J., Weidmann, N.B.: Using machine-coded event data for the micro-level study of political violence. *Research & Politics* **1**(2), 2053168014539924 (2014). <https://doi.org/10.1177/2053168014539924>, <https://doi.org/10.1177/2053168014539924>

8. Hoffmann, M., Santos, F.G., Neumayer, C., Mercea, D.: Lifting the veil on the use of big data news repositories: A documentation and critical discussion of a protest event analysis. *Communication Methods and Measures* **16**(4), 283–302 (2022). <https://doi.org/10.1080/19312458.2022.2128099>, <https://doi.org/10.1080/19312458.2022.2128099>
9. McAuliffe, M., Triandafyllidou, A. (eds.): *World Migration Report 2022*. International Organization for Migration (2021)
10. Napierała, J., Hilton, J., Forster, J.J., Carammia, M., Bijak, J.: Toward an early warning system for monitoring asylum-related migration flows in Europe. *International Migration Review* **56**(1), 33–62 (2022). <https://doi.org/10.1177/01979183211035736>
11. Philip A. Schrod: *Conflict and Mediation Event Observations Event and Actor Codebook (CAMEO)*. <http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf> (2012)
12. The GDELT Project: GDELT 2.0: Our Global World in Realtime. <https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/> (2015)
13. The GDELT Project: New November 2021 GKG 2.0 Themes Lookup. <https://blog.gdeltproject.org/new-november-2021-gkg-2-0-themes-lookup/> (2021)
14. The GDELT Project, GCAM: Introducing the Global Content Analysis Measures (GCAM). <https://blog.gdeltproject.org/introducing-the-global-content-analysis-measures-gcam/> (2014)
15. The GDELT Project, GKG: Introducing GKG 2.0 – The Next Generation of the GDELT Global Knowledge Graph. <https://blog.gdeltproject.org/introducing-gkg-2-0-the-next-generation-of-the-gdelt-global-knowledge-graph/> (2014)
16. Tilly, S., Ebner, M., Livan, G.: Macroeconomic forecasting through news, emotions and narrative. *Expert Systems with Applications* **175**, 114760 (2021). <https://doi.org/10.1016/j.eswa.2021.114760>
17. UNHCR: Figures at a Glance. <https://www.unhcr.org/en-us/figures-at-a-glance.html> (Jun 2022)
18. Vargo, C.J., Guo, L., Amazeen, M.A.: The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society* **20**(5), 2028–2049 (2018). <https://doi.org/10.1177/1461444817712086>
19. Voukelatou, V., Miliou, I., Giannotti, F., Pappalardo, L.: Understanding peace through the world news. *EPJ Data Science* **11**(2) (2022). <https://doi.org/10.1140/epjds/s13688-022-00315-z>
20. Weisser, R.A.: A near-real-time analysis of societal responses to Ukrainian refugee migration in Europe. *International Migration* (Nov 2022). <https://doi.org/10.1111/imig.13071>
21. Yesilbas, V., Padilla, J.J., Frydenlund, E.: An analysis of global news coverage of refugees using a big data approach. In: *Proceedings of the 14th International Conference on Social, Cultural, and Behavioral Modeling (SBP-BRiMS)*. pp. 111–120 (2021). https://doi.org/10.1007/978-3-030-80387-2_11
22. Yonamine, J.E.: A nuanced study of political conflict using the Global Datasets of Events Location and Tone (GDELT) dataset. Ph.D. thesis, The Pennsylvania State University (Aug 2013)