Tracking COVID-19 protest events in the United States. Shared Task 2: Event Database Replication, CASE 2022

Vanni Zavarella
University of Cagliari
v.zavarella@unica.it

Hristo Tanev
European Commission
hristo.tanev@ec.europa.eu

Ali Hürriyetoğlu
KNAW Humanities Cluster DHLab
ali.hurriyetoglu@dh.huc.knaw.nl

Peratham Wiriyathammabhum
peratham.bkk@gmail.com

Bertrand De Longueville
European Commission
bertrand.de-longueville@ec.europa.eu

Abstract

The goal of Shared Task 2 is evaluating state-of-the-art event detection systems by comparing the spatio-temporal distribution of the events they detect with existing event databases.

The task focuses on some usability requirements of event detection systems in real world scenarios. Namely, it aims to measure the ability of such a system to: (i) detect socio-political event mentions in news and social media, (ii) properly find their geographical locations, (iii) de-duplicate reports extracted from multiple sources referring to the same actual event. Building an annotated corpus for training and evaluating jointly these sub-tasks is highly time consuming. One possible way to indirectly evaluate a system's output without an annotated corpus available is to measure its correlation with human-curated event data sets.

In the last three years, the COVID-19 pandemic became motivation for restrictions and anti-pandemic measures on a world scale. This has triggered a wave of reactions and citizen actions in many countries. Shared Task 2 challenges participants to identify COVID-19 related protest actions from large unstructured data sources both from mainstream and social media. We assess each system's ability to model the evolution of protest events both temporally and spatially by using a number of correlation metrics with respect to a comprehensive and validated data set of COVID-related protest events (Raleigh et al., 2010).

1 Introduction

State-of-the-art evaluation methods for event detection are based on manually coded corpora with annotated document and sub-document units, including annotation of syntactic fragments, such as event reporting verbal phrases, as well as entities having specific semantic roles, such as *victim, perpetrator, weapons*, etc., see (Hürriyetoğlu et al., 2021) and (Atkinson et al., 2017) among the others. While this type of benchmarks provide accurate means for measuring the performance of event detection approaches, their development implies significant efforts: many person-hours of annotations by journalists or linguists, which make such annotated corpora limited in number and size and generally developed for the English language only, with a few exceptions (Hürriyetoğlu et al., 2021). Moreover, such evaluation methods do not assess the overall usability of machine-coded event data sets for micro-level modelling of social processes. Also, in the domain of socio-political and armed conflicts, spatio-temporal analysis has become standard and state-of-the-art evaluation methods come short in evaluating exhaustively the spatial and temporal aspects of event detection systems.

Extracting spatio-temporal information from online text sources has developed in the late 2000's, with the advent of the so-called 'Web 2.0' and Social Networks (Pultar et al., 2008), (De Longueville et al., 2009). Since then, applications have been developed in fields as diverse as disaster management (De Longueville et al., 2010), traffic monitoring (D'Andrea et al., 2015), or fight against crime (Kounadi et al., 2015). Detecting socio-political events (and in particular, protests) emerged as an important use case (Zhang, 2019), as many applications in this field need to rely on comprehensive, timely and high-quality data that is often not available (e.g. high quality commercial data is produced on a weekly, or even monthly basis, while applications need near-real-time data). This is a gap that CASE workshops, and this shared task in particular, are aiming to address.

The dynamics of the COVID-19 protests and

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 209 - 216

their varied media coverage by news outlets and social media makes it a particularly relevant use case for assessing the capability of automated event extraction systems to analyse socio-political processes. The database replicability Shared Task 2 aims at doing so by challenging event extraction systems to extract a collection of protest events from two heterogeneous text collections (i.e., news and social media). The task's evaluation is done by measuring a number of spatio-temporal correlation coefficients against a gold standard data set of protest incidents, provided by the the Armed Conflict Location and Event Data (ACLED) project (Raleigh et al., 2010).

This task is the second in a series of shared tasks at the CASE 2022 workshop (Hürriyetoğlu et al., 2022b). The first task is concerned with protest news detection at multiple text resolutions (e.g., the document and sentence level) and in multiple languages: English, Hindi, Portuguese, and Spanish (Hürriyetoğlu et al., 2021, 2022a). Task 3 is about detecting event causality in a corpus of sentence pairs that have been annotated with labels on whether there is a causal relations or not between them (Tan et al., 2022a,b).

Teams which participated in Task 1 were invited to participate in this second task. This is an evaluation only task, where all models are (i) trained on the data provided in Task 1, (ii) applied to raw news and social media data, specifically gathered for the task (i.e, news collection crawled from the Web from various news sources, as well as Twitter data), and (iii) evaluated on a manually curated, COVID-19 protest event list, gathered from the Web page of the ACLED project (Raleigh et al., 2010).

2 Related Work

Some recent studies show that performance measures such as precision, recall, and F1 are limited in their capacity to asses the efficiency of an NLP system (Derczynski, 2016; Yacouby and Axman, 2020). Moreover, evaluating a system on detecting socio-political events from text requires additional metrics such as spatio-temporal correlation of the system output and the actual distribution of the events (Wang et al., 2016; Althaus et al., 2021).

In a detailed study Cook and Weidmann (2019) demonstrates the usefulness of disaggregating event reports when considering data from event coding. Several approaches deal with assessing the correlation of automatically generated event data sets with gold standards based on disaggregated event counts, see example Ward et al. (2013) and Schrodt and Analytics (2015) among the others.

Hammond and Weidmann (2014) applied disaggregation of events across PRIO-GRID geographical cells (Tollefsen et al., 2012) to assess the spatio-temporal pattern of conflicts in the Global Database of Events, Language and Tone (GDELT) (Leetaru and Schrodt, 2013). In a later work Zavarella et al. (2020) adapted the aforementioned approach to administrative units for measuring the impact of event de-duplication on increasing correlation with ACLED event data sets.

3 Data

The goal of this task is to evaluate the performance of automatic event detection systems on modeling the spatial and temporal pattern of a social protest movement. We evaluate the capability of participant systems to reproduce a manually curated COVID-19 related protest event data set, by detecting protest event reports, enriched with location and date attributes, from a news corpus collection, a Twitter collection (both pre-filtered for COVID-19 topic occurrence) and from the union of the two.

3.1 Training Data

As a usability analysis, no training data were provided for this Task. Namely, the event definition applied for coding the reference event data set is the same as the one adopted for Shared Task 1 (Hürriyetoğlu et al., 2021) and any data utilized for Task 1 and Task 2, such as the one from Hürriyetoğlu et al. (2021); Duruşan et al. (2022); Yörük et al. (2021), or any additional data could be used to build a system/model run on the input data.

3.2 Input Data

We provide three collections of input data:

- an English language news corpus comprising a large selection of COVID-related articles from US news sources;
- an English language tweet collection comprising daily samples of COVID-related tweets with some geographical metadata referring to U.S.;
- a Spanish language tweet collection comprising daily samples of COVID-related tweets

with some geographical metadata referring to U.S.

News Collection The news corpus used in this Task is a collection of articles in English language spanning the time range July 27, 2020 through October 26,2020 from a large set of news sources from U.S. We used public APIs when available and scraped the newspaper web pages otherwise. For example, we used the New York Times Archive API [1]. The articles are filtered by checking the occurrence of keywords ["covid","coronavirus"] in the top two sentences of the articles. Overall the collection contains around 122k articles. We harmonized the news item metadata from the different collections so as to have the attributes: Publication Date of the article, Title and a Snippet from the article text, comprising the 2 lead sentences.

Twitter The corpus used in this Task is based on a large-size multilingual collection of tweets sampled from the Twitter public streaming API using the set of keywords ["*COVID19*", "*CoronavirusPandemic*", "*COVID-19*", "*2019nCoV*", "*CoronaOutbreak*", "*coronavirus*", "*WuhanVirus*"], described in (Banda et al., 2021). The source data of this collection, together with documentation on how to process the data, can be found on https://github.com/thepanacealab/covid19_twitter.

We used the clean version of this dataset that was already filtered for retweets. The collection of tweets is language tagged since July 27 2020. We further filter the data from July 27, 2020 through October 26, 2020 and produce two monolingual tweet collections for English and Spanish. Namely, in order to restrict the sample to content from the US context, we filter for tweets which have a *Place* metadata with *Place*'s *country_code="us"* or (if *Place* is None) with a *User* location specified as one of the US States. For each day, we filter up to reaching a sampling cap ratio of 0.1 and 0.5 of the original tweet collections for English and Spanish, respectively. The overall size of the tweet collections are about 2.8M and 46k for English and Spanish, respectively, with an average of 30k and 503 tweets per day. We distributed the numeric tweet ids and participants were allowed to process any of the tweet's meta-data for their system runs.

[1] https://developer.nytimes.com/get-started

3.3 Gold Standard Data

We challenge the participant systems to reproduce a Gold Standard data set from the ACLED project's COVID-19 Disorder Tracker[2], comprising curated disorder events directly related to the coronavirus pandemic.

These include: a.targeting of healthcare workers responding to the coronavirus, b.violent mobs attacking individuals arbitrarily viewed as linked to the coronavirus and c. demonstrations against response measures to coronavirus (government's lock-downs, etc). On the other hand, changes in already existing demonstration patterns as a result of coronavirus-related restrictions, or disorder events driven by already existing armed or political group capitalizing on the coronavirus-induced instability are not included in the data set. From the whole data set, we select events tagged with ACLED types *Protest* and *Riot* and with a US country code location, for the time range from July 27, 2020 through October 26, 2020, resulting with a set of 1449 events, with event date, city, state, country-level information and geographical coordinates.

Notice that while ACLED data come with both hierarchical, string-like location information (i.e. place names at different administrative levels) and coordinate pairs, for the sake of consistency with system output results we re-processed string-like location descriptions of Gold Standard events using the method described in 4.1 and re-generated event coordinate pairs before joining with PRIO-GRID shapefiles.

The U.S. map in Figure 1 shows the spatial distribution of these events (blue dots).

4 Evaluation

System performance is evaluated by computing correlation coefficients on event counts aggregated on cell-days, using uniform grid cells of approximately 55 kilometers sides from the PRIO-GRID data set (Tollefsen et al., 2012). We use these analytical measures as a proxy to the spatio-temporal pattern of the coronavirus-related protest events.

4.1 Data Normalization

In order to be joined with PRIO-GRID shapefiles, string-like location information of system output data had to be normalized to coordinate pairs. To do this we used the OpenStreetMap Nominatim

[2] https://acleddata.com/analysis/covid-19-disorder-tracker/

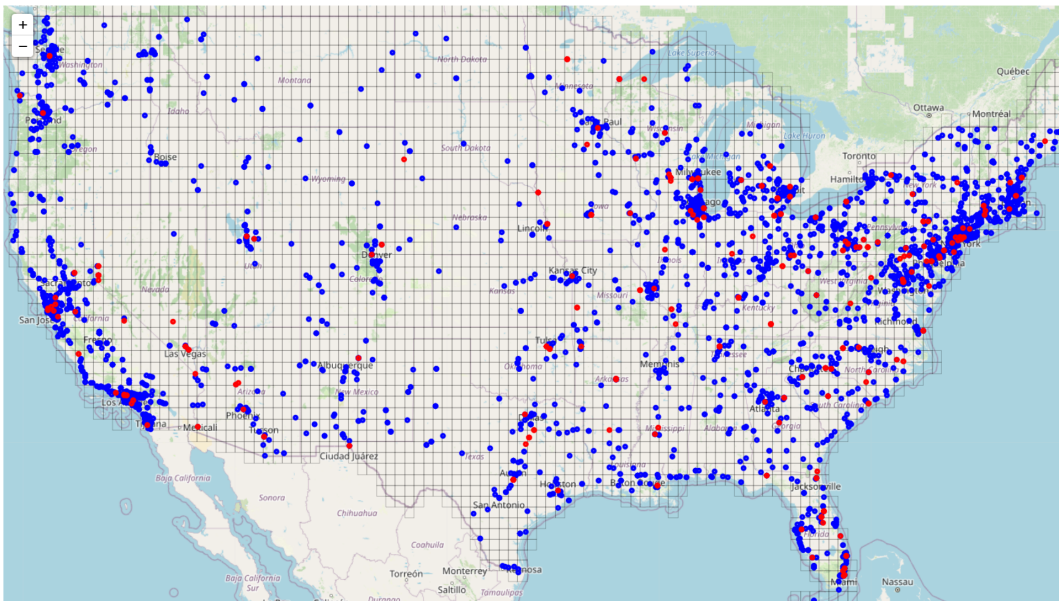

Figure 1: The geo-referenced COVID-19 protest event records from ACLED (blue dots) and the events automatically detected by the *Classbases* system (red dots), overlaid with the PRIO-GRID cells over the US.

search API[3]. For structured location name representations (i.e., *city*, *state*, *country*) we used a parametric search: if this fails, we back off to free-form query strings.

We note that geographical coordinate conversion from Nominatim places the event at the geographical centroid of the polygon of the assigned administrative unit. In our evaluation, we discarded the system output event records with no source location information or whose string-like location attribute returned *Null* results through the Nominatim API.

4.2 Metrics

We use the cell-days counts for two different types of analysis: the correlation with the total daily "protest cell" counts (i.e., time trends alone) and the event counts for each cell-day (i.e., spatial and temporal trends together).

Temporal Trends The first analysis only considers the correlation between the total number of "activated" cells (i.e., for which at least one Protest event was recorded), in the system output and the Gold Standard data set.

This time series analysis is sufficient to estimate how well the automatic system captures the development in time of the protest movement, without considering the geo-location accuracy. So, it evaluates on the task of detecting or not an event in the document collection.

Spatial and Temporal Trends To this purpose, we also measure the correlation coefficients on the absolute event counts with respect to Gold Standard, over each single cell-day. In this way also the geolocation capabilities of the system are considered.

For both analyses, we use two types of correlation coefficients to assess variable's relationship: Pearson coefficient r and Spearman's rank correlation coefficient ρ. Moreover, we used Root Mean Squared Error (RMSE) to measure the absolute value of the error on estimating cell/event counts from the Gold Standard.

4.3 Team Systems

Only one team participated in this edition of the Shared Task: *Classbases*. We briefly describe the system below and ask the reader to refer to their system paper for additional details (Wiriyathammabhum, 2022).

Classbases The *Classbases* system used the trained XLM-RoBERTa large model from subtask1 to classify the news using a concatenation of its news title and news abstract to guess whether it contains any protest events or not. If the classifier outputs positive (logits were thresholded at 0.9), we ran a SpaCy named entity recognizer (Honnibal et al., 2020) on the textual concatenation to get spans with location tags ('GPE'). Then, those spans were

[3] https://nominatim.org/release-docs/develop/api/Search/#parameters

	Data	r	ρ	RMSE
Classbases	News	-0.330	-0.331	193.60

Table 1: Correlation coefficients and error rates for daily protest cell counts: r represents Pearson correlation coefficient, ρ is Spearman's rank correlation coefficient, and RMSE is the Root Mean Squared Error computed on day-cell units.

concatenated into a query string which we used a geocoder library2 to geocode using the Bing Maps REST Services API3. We used the provided dates from the date column as outputs given the filtered ids. Finally, we outputted a row for each filtered id containing five tuples, which are the id, the date, the city, the region or state, and the country.

5 Results

Table 1 shows the Pearson r, Spearman correlation coefficient ρ, and Root Mean Squared Error (RMSE) for the total daily protest cell counts over the 92 days target time range of the only participant system,*Classbases*, run on the news data (denoted as *Classbases_new_1* in the plot.

Here, the correlations are between the total number of cells per day where the system found an event vs. the number of cells where at least one event occurred according to the Gold Standard.

The figures show no correlation between the automatically detected conflict cells and the gold standard over time. This is evident from Figure 2, where we plot the time series of total daily protest cells of the participant system against Gold Standard. We see the system evaluated on news data failing to pick up both temporal variation (i.e., the gradually declining weekly picks of protests from early August through October) and the overall magnitude of the protest movement (e.g., it detects a maximum of less then 10 protest cells per day).

While this correlation analysis is overall more tolerant to errors in geocoding[4], a more in-depth error analysis showed that geocoding inaccuracy caused: a. several detected events to wrongly activate the same cells in system output, causing the geographical spread to be significantly lower than Gold Standard; b. some highly recurrent place names to be wrongly resolved to multiple homonym locations, activating additional cells.

Table 2 reports Pearson r, Spearman correlation coefficient ρ, and Root Mean Squared Error (RMSE) over cell-day event counts of the *Classbases* system with respect to Gold Standard for the 92 days time range

Here the variables range over the whole set of PRIO-GRID cells included in the US territory and, thus, show the correlation of event numbers across geo-cells, thus better evaluating the system's fine-grained geolocation capabilities. As expected, no significant correlation with Gold Standard is found here either.

A more lenient representation of the agreement with Gold Standard is shown in Table 3. Here we report the confusion matrix between grid cells that Gold Standard and system runs code as experiencing at least a protest event. It can be observed that only few of the cells classified as Protest by Gold Standard are detected by the automatic system, which on the other hand incorrectly classified as Protest several additional cells.

6 Conclusions

The goal of the "Covid protest events" Shared Task was to explore novel performance evaluations of pre-trained event detection systems. These systems are applied to large noisy, heterogeneous text data sets (i.e., news articles and social media data) related to a specific protest movement or, as in this case, a wave of protests induced by the coronavirus crisis. Thus, the systems are being evaluated out-of-domain in terms of both data type (i.e., the systems are trained on news data and evaluated on both news and social media) and protest movement context (i.e., the training data are not necessarily related to covid-19 pandemic). Systems are evaluated on their ability to identify both events across time as well as their distribution across space. This evaluation scenario proved challenging for the system participating in the shared task, confirming the finding from the previous edition(Giorgi et al., 2021). A major problem, as shown on Table 3, is the systems' low recall.

The low recall at this years shared task may be due to the pre-filtering of the news data for the presence of covid-19 mentions. Differently than for an organized protest movement (like Black Lives

	Data	r	ρ	RMSE
**Classbases*	News	0.0247	0.0342	0.0101

Table 2: Correlation coefficients and error rates for *cell-day* event counts of the participant systems with respect to Gold Standard.

[4]Indeed, as long as the events are located in U.S., a systematic misplacement of the events might not potentially affect its geographical 'spread' in terms of number of activated cells.

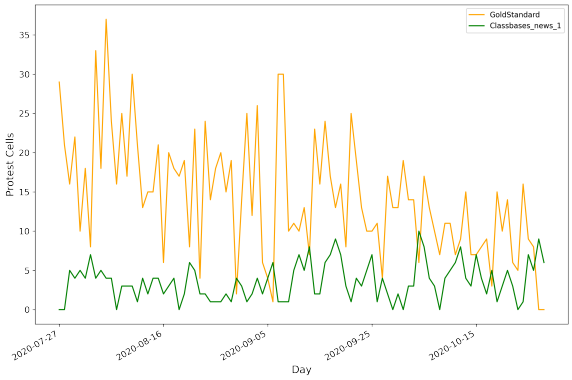

Figure 2: Time series of total daily protest cells from the Gold Standard (in orange), against the *Classbases* system run on news data.

		Gold Standard		Precision	Recall	F1
		1	0			
Classbases	1	24	312	7.14	1.76	2.83
	0	1333	478765			

Table 3: Confusion matrix of grid cells experiencing at least one Protest event (true) versus inactive cells (false), for the Gold Standard and participant systems. Given the high negative class imbalance of the data, we report Precision,Recall figures for the positive class only.

Matter), inferring a relationship of single protest events to the pandemic might not be trivial and thus explicitly stated in the protest news report: therefore, filtering for covid-19 keywords might remove relevant protest reports. However, absolute low recall does not necessarily affect correlation measures as much as inaccurate geocoding of the detected events, as shown.

Overall, this year's edition of the Task was compromised by the low attendance and it is not possible to draw further significant conclusions. We therefore decided to re-open the evaluation window open and welcome further system run submissions. Researchers interested to have their models run and evaluated on the input data provided can check the GitHub `https://github.com/zavavan/case2022_task2` or contact the authors.

Acknowledgments

The author from Koc University was funded by the European Research Council (ERC) Starting Grant 714868 awarded to Dr. Erdem Yörük for his project Emerging Welfare.

References

Scott Althaus, Buddy Peyton, and Dan Shalmon. 2021. A total error approach for validating event data. *American Behavioral Scientist*, 3(2).

Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. 2017. On the creation of a security-related event corpus. In *Proceedings of the Events and Stories in the News Workshop*, pages 59–65.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324.

Scott J Cook and Nils B Weidmann. 2019. Lost in aggregation: Improving event analysis with report-level data. *American Journal of Political Science*, 63(1):250–264.

Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni. 2015. Real-time detection of traffic from twitter stream analysis. *IEEE transactions on intelligent transportation systems*, 16(4):2269–2283.

Bertrand De Longueville, Gianluca Luraschi, Paul Smits, Stephen Peedell, and Tom De Groeve. 2010. Citizens as sensors for natural hazards: A vgi integration workflow. *Geomatica*, 64(1):41–59.

Bertrand De Longueville, Robin S Smith, and Gianluca Luraschi. 2009. " omg, from here, i can see the

flames!" a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks*, pages 73–80.

Leon Derczynski. 2016. Complementarity, F-score, and NLP evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266, Portorož, Slovenia. European Language Resources Association (ELRA).

Fırat Duruşan, Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Çağrı Yoltar, Burak Gürel, and Alvaro Comin. 2022. Global contentious politics database (glocon) annotation manuals.

Salvatore Giorgi, Vanni Zavarella, Hristo Tanev, Nicolas Stefanovitch, Sy Hwang, Hansi Hettiarachchi, Tharindu Ranasinghe, Vivek Kalyan, Paul Tan, Shaun Tan, Martin Andrews, Tiancheng Hu, Niklas Stoehr, Francesco Ignazio Re, Daniel Vegh, Dennis Atzenhofer, Brenda Curtis, and Ali Hürriyetoğlu. 2021. Discovering black lives matter events in the United States: Shared task 3, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 218–227, Online. Association for Computational Linguistics.

Jesse Hammond and Nils B Weidmann. 2014. Using machine-coded event data for the micro-level study of political violence. *Research & Politics*, 1(2):2053168014539924.

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Selçuk Gürel, Benjamin Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022a. Extended Multilingual protest news detection - Shared Task 1, CASE 2021 and 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Osman Mutlu, Farhana Ferdousi Liza, Erdem Yörük, Ritesh Kumar, and Shyam Ratan. 2021. Multilingual protest news detection - Shared Task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Reyyan Yeniterzi, Osman Mutlu, and Erdem Yörük. 2022b. Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022): Workshop and Shared Task Report. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021. Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction. *Data Intelligence*, 3(2):308–335.

Ourania Kounadi, Thomas J Lampoltshammer, Elizabeth Groff, Izabela Sitko, and Michael Leitner. 2015. Exploring twitter to analyze the public's reaction patterns to recently reported homicides in london. *PloS one*, 10(3):e0121848.

Kalev Leetaru and Philip A Schrodt. 2013. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Edward Pultar, Martin Raubal, and Michael F Goodchild. 2008. Gedmwa: Geospatial exploratory data mining web agent. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pages 1–4.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660.

Philip A Schrodt and Parus Analytics. 2015. Comparing methods for generating large scale political event data sets. In *Text as Data meetings, New York University, 16–17, 2015*, pages 1–32.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Hansi Hettiarachchi, Tadashi Nomoto, Onur Uca, and Farhana Ferdousi Liza. 2022a. Event Causality Identification with Causal News Corpus - Shared Task 3, CASE 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022b. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Andreas Forø Tollefsen, Håvard Strand, and Halvard Buhaug. 2012. Prio-grid: A unified spatial data structure. *Journal of Peace Research*, 49(2):363–374.

Wei Wang, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. 2016. Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1503.

Michael D Ward, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013. Comparing gdelt and icews event data. *Event Data Analysis*, 21(1):267–297.

Peratham Wiriyathammabhum. 2022. ClassBases at CASE-2022 Multilingual Protest Event Detection Tasks: Multilingual Protest News Detection and Automatically Replicating Manually Created Event Datasets. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, online. Association for Computational Linguistics (ACL).

Reda Yacouby and Dustin Axman. 2020. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 79–91, Online. Association for Computational Linguistics.

Erdem Yörük, Ali Hürriyetoğlu, Fırat Duruşan, and Çağrı Yoltar. 2021. Random sampling in corpus design: Cross-context generalizability in automated multicountry protest event collection. *American Behavioral Scientist*, 0(0):00027642211021630.

Vanni Zavarella, Jakub Piskorski, Camelia Ignat, Hristo Tanev, and Martin Atkinson. 2020. Mastering the media hype: Methods for deduplication of conflict events from news reports. In *Proceedings of AI4Narratives — Workshop on Artificial Intelligence for Narratives*.

Shuo Zhang. 2019. Data mining Mandarin tone contour shapes. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 144–153, Florence, Italy. Association for Computational Linguistics.