Post-Soviet Affairs

ISSN: 1060-586X (Print) 1938-2855 (Online) Journal homepage: www.tandfonline.com/journals/rpsa20

# Collecting protest event data using natural language processing models

Bogdan Mamaev

**To cite this article:** Bogdan Mamaev (11 Dec 2025): Collecting protest event data using natural language processing models, Post-Soviet Affairs, DOI: 10.1080/1060586X.2025.2600874

**To link to this article:** https://doi.org/10.1080/1060586X.2025.2600874

POST-SOVIET AFFAIRS
https://doi.org/10.1080/1060586X.2025.2600874

Routledge
Taylor & Francis Group

RESEARCH ARTICLE



# Collecting protest event data using natural language processing models

Bogdan Mamaev

Australian Internet Observatory, School of Humanities & Social Sciences, Deakin University, Melbourne, Australia

**ABSTRACT**

This article presents the Russian Contentious Events Dataset (RCED), a comprehensive dataset of contentious events in Russia from 2010 to 2023. It addresses the challenge of collecting protest event analysis (PEA) data by using social media and natural language processing (NLP) models to automatically identify and analyze reports of such events. The article details a methodological workflow that includes event classification, named entity recognition, duplicate removal, and post-processing. Analysis of the generated dataset reveals longitudinal protest trends in Russia. Using summaries of the original tweets, the paper demonstrates how event classification and spatial clustering can be used to analyze contention at both federal and regional levels, identifying significant variations across the country. This study shows that modern machine learning and language models can automate and scale PEA, improving the data collection process while reducing resource requirements. The article contributes to the literature on the automation of protest data collection and emphasizes the need for further research into how advances in NLP can be applied in this field.

**ARTICLE HISTORY**
Received 5 April 2025
Accepted 25 November 2025

**KEYWORDS**
Contentious events analysis; protest event dataset; protests in Russia; contentious action in Russia; RCED

Protest event analysis (PEA) is a data analysis technique used to quantify and compare levels of contention across time and regions (Dollbaum 2021; Koopmans and Statham 1999; Olzak 1989). As a type of quantitative content analysis, PEA helps link social and political processes to protest, making it a crucial tool in the study of contentious politics. However, the application of PEA presents challenges, as researchers must often synthesize scarce data from diverse sources varying in detail and reliability. Historically, the field has relied heavily on newspapers (Earl et al. 2004) and online publications (Weidmann and Rød 2019b), requiring extensive manual data collection and coding efforts by numerous contributors (Croicu and Weidmann 2015; Kriesi et al. 1995; Lankina and Tertytchnaya 2019). Linking contention to societal and political developments (Fisher et al. 2019), PEA remains a costly and time-consuming process, often resulting in data limitations related to reliability, verifiability, and aggregation (Weidmann and Rød 2019b).

**CONTACT** Bogdan Mamaev bodgan.mamaev@deakin.edu.au Australian Internet Observatory, School of Humanities & Social Sciences, Deakin University, 221 Burwood Highway, Burwood, VIC 3125, Melbourne, Australia
Supplemental data for this article can be accessed online at https://doi.org/10.1080/1060586X.2025.2600874

These challenges are amplified in authoritarian regimes. Limited free media, restricted international coverage, political persecution, and information control create additional biases and obstacles, complicating dataset creation and the methodology required to acquire such data (Dollbaum 2021). One such case is the authoritarian regime in Russia, which actively uses disinformation (Mejias and Vokuev 2017) and maintains state control over traditional (Dovbysh and Mukhametov 2020; Lipman 2009) and social media (Poupin 2021). This is coupled with repression of journalists and activists (Lipman 2016) alongside state and self-censorship by reporters, media, and users (Lamberova and Sonin 2023; Roudakova 2017; Schimpfössl and Yablokov 2020). These limitations increase reliance on activist and reporter networks (Lankina and Tertytchnaya 2019) or online and regional media (Bizyukov and Dollbaum 2021). The regime's actions likely cause event underreporting due to repressive threats and the difficulty of obtaining reliable data.

While analytical quality depends on the dataset's quality (Dollbaum 2021), underreporting of protests in event datasets is inevitable (Weidmann and Rød 2019b). This is worsened by overreliance on English-language and international media, which often have limited coverage of smaller or non-urban protests. To mitigate this, scholars use computational approaches for more efficient data collection and analysis, applying machine learning (ML), including neural networks and natural language processing (NLP) (Lorenzini et al. 2022; Zhang and Pan 2019). Some integrate computation with manual coding (Lorenzini et al. 2022; Weidmann and Rød 2019a), while others develop fully automated workflows (Hanna 2017; Zhang and Pan 2019). These advancements aim for deeper insights into contention by reducing reliance on human coders. Moreover, increasing PEA efficiency through additional sources like social media and images, beyond news reports, is a significant step towards understanding contention across regime types.

Developments in NLP, along with other computational techniques, offer various methods to facilitate PEA and the generation of high-quality datasets. The recent prominence of transformer-based ML models and other large language models (LLMs) offers potential solutions. These tools are useful for data collection, automated pre- and post-processing, coding, duplicate removal, and classification based on specific attributes such as protest type or protester claims. However, the field of contentious politics has not kept pace with these developments. There is a general lack of awareness of these tools in the literature, indicating a slow rate of incorporating these interdisciplinary advancements into research. Despite the sparse application of some computational methods, comparative analysis evaluating their efficiency and potential to complement or replace existing approaches remains notably absent. Insufficient research on how computational techniques can assist in preparing protest datasets leaves their performance and potential for integration with existing PEA methods unclear, particularly when combining multiple data sources.

This paper examines how contentious action evolved in Russia during the 2010s amid an increasingly repressive environment that culminated in 2022–2023. In doing so, it advances the study of authoritarian regimes by proposing a computational workflow for PEA. To demonstrate how modern computational methods and social media data can be used to study contention, this research introduces a comprehensive contentious events dataset and the methodological workflow used to create it. The resulting Russian Contentious Events Dataset (RCED) draws on descriptions from thousands of Twitter

accounts to address potential underreporting and bias. While many event datasets focus on specific categories or issues, RCED encompasses a broader range of contentious events. Finally, the dataset is compared with other established PEA data sources, including GDELT, ICEWS, PLOVER, MMAD, and LaRUPED.

The methodological workflow in this paper addresses the limitations of manual coding by presenting an automated approach to protest data collection, thereby reducing the associated resource burden. Drawing on recent advances in ML and NLP, this study evaluates the suitability of these models as potential substitutes for manual coding in PEA. The methodology employed for RCED utilizes Transformer models, network analysis, contextual deduplication, named entity recognition (NER), translation, and summarization. This research also contributes to the growing field of automated PEA and the use of social media data in protest analysis.

The following sections discuss PEA and its most recent developments, particularly the use of automated and manual event classification methods. This research further proposes a context-specific approach to data collection, detailing the employed tools and techniques while evaluating their suitability and performance step by step. A comparison is then made between the resulting dataset and other existing datasets. Using RCED, the analysis subsequently explores how contentious action evolved in Russia from 2010 to 2023, utilizing topic modeling, clustering, and multidimensional scaling (MDS). In conclusion, the discussion addresses limitations and avenues for future research.

## Challenges and opportunities in protest event data collection

### *The evolution of PEA*

The development of PEA coincided with advances in social movement studies. Bridging interdisciplinary advances in political sociology and social and cultural psychology led to theoretical and methodological developments in social movement research, including the application of event analysis (Oliver, Cadena-Roa, and Strawn 2003). As a largely quantitative technique, its application grew in the 1980s and 2000s. Numerous studies focused on quantifying protest events, measuring changes in their trends, and testing theoretical hypotheses related to political opportunity theory (Hutter 2014; Koopmans and Statham 1999; Oliver, Cadena-Roa, and Strawn 2003). Such event-centric analysis allows for a deeper understanding of mobilization dynamics, explaining why certain events occur while others do not. It helps identify predictors for particular categories of events, measure time dynamics and causality, control for independent variables, and enable operationalization and empirical testing, particularly in longitudinal studies (Oliver, Cadena-Roa, and Strawn 2003, 220–222).

Major works in this area include McAdam's (1999) analysis of the rise and decline of the US civil rights movement from 1930 to 1970 using coded newspaper data. Earl, Soule, and McCarthy (2003) employed PEA to explain factors predicting police presence and strategy deployment during protests in New York from 1968 to 1973. Lankina and Skovoroda (2017) explored the relationship between electoral fraud and protest in 95,415 voting precincts during the 2012 presidential elections in Russia, discovering that fraud was a contributing factor to post-electoral regional protests. In another example, using monthly occurrence and participation numbers during

protests in 30 European countries, Kriesi and Oana (2022) observed that restrictions placed during COVID-19 lockdowns significantly impacted protest occurrence, shifting the focus of protest events to issues directly related to COVID-19, such as lockdown restrictions. King and Soule (2007) used quantitative corporate activist protest event data from 1962 to 1990 in the US and applied social movements theory to investigate how protests impacted stock prices.

The increasing diversity of PEA applications has led to further efforts to develop data collection techniques that reduce biases and address limitations in traditional methods. Advancements in the field and new data technologies have led to the development of innovative methods and datasets aimed at providing sufficient, high-quality data for PEA while reducing the workload associated with data collection.

## *Newspapers as a source of data*

Traditionally, PEA relied on newspaper articles as the main source of quantitative data or event catalogs. Due to the scarcity of official reports, researchers often drew upon archives, periodicals, interviews, observations, and other forms of reports (Koopmans and Statham 1999; McAdam 1999; Olzak 1992; Tarrow 1989). Researchers coded thousands of reports from newspapers and other sources, utilizing either newspaper indices (Benson and Saxton 2010) or full articles (Johnson, Schreiner, and Agnone 2016) for further analysis. However, the utility of newspapers as a source for accurately representing reality and accounting for contentious events has been questioned. The limitations include limited coverage, inherent biases due to political or other preferences, and the issue of underreporting (Hutter 2019; Strawn 2008). These limitations require additional methods of analysis and testing to ensure the validity and reliability of findings. However, as Hutter (2019) points out, this issue has not always been addressed in research.

Furthermore, the scarcity of data often makes it difficult to assess the presence and extent of bias in existing sources. Mueller (1997) conceptualized this problem as the issue of not knowing how specific media sources select what data to report, leading to the acceptance of inherent biases at face value. This becomes especially important when studying the dynamics of particular events and their repertoires rather than simple listings of events. Earl et al. (2004) provided a comprehensive overview of potential sources of bias in newspaper data, including production processes, reporter routines, and prevailing social concerns. Davenport (2009, 31) points out that bias lies in the "incomplete representation of conflict behaviour reported in sources," while others argue that the use of multiple sources of protest data should be prioritized over the use of single-source data (Nam 2006).

## *Online reports and computational analysis*

The use of the Internet and computational methods has further advanced PEA. Researchers have been gathering data from online resources, which makes cataloging and analysis more efficient and less resource intensive. While protest event catalogs based on media reports still prevail (Weidmann and Rød 2019b), manual data collection has been increasingly complemented by web scraping and data extraction from online sources, including websites and electronic versions of traditional media outlets, leading

to greater efficiency in data gathering and processing (Andretta and Pavan 2018). For example, Lankina and Tertytchnaya (2019) utilized human coders to catalog events published by a network of activists and reporters around Russia for their dataset on Russian protests.

As ML tools gain prominence, more researchers have begun relying on them for tasks such as event coding, classification, and filtering. Recent works often employ a human-in-the-loop approach, where automated tools are supervised by human coders or handle specific tasks, leading to improved accuracy and reliability. For instance, Borbáth and Hutter (2021) and Kriesi and Oana (2022) employed a dataset generated using machine learning for initial content analysis before human coders conducted a more detailed analysis of reports from European newswires. Similarly, Weidmann and Rød (2019b) generated a report-level and event-level dataset on protests in autocracies by using machine learning to identify and pre-filter potential reports from multiple English-language sources before manual coding.

The development of ML and NLP has sparked discussions on how these methods can be applied to protest event data. To address the problem of insufficient data, Wüest, Rothenhäusler, and Hutter (2013) suggested using topic and word space models together with named entity recognition (NER) to code a dataset of protest events. One of the most commendable efforts to date is the work of Zhang and Pan (2019). They implemented recurrent and convolutional networks on image and textual data to create a fully automated approach for data collection on the heavily censored Chinese social platform Weibo, relying on user reports. They also noted one of the main advantages of using social media as a tool for data collection: it empowers every user to broadcast and report on events, potentially surpassing the reach and scope of traditional media, particularly in authoritarian regimes.

Another important effort was the Global Database of Events, Language, and Tone (GDELT), which employs tens of thousands of multilingual sources and websites and constantly extracts information from them (Leetaru and Schrodt 2013).[1] GDELT has been applied in PEA research (Hoffmann et al. 2022) and has proven useful in some analyses of protests. However, it has also been criticized for overreporting specific events and including false positives, raising concerns about validity, transparency, and reliability. These issues, such as the repeated reporting of the same event or the inclusion of non–protest-related articles, have implications for analysis that should be carefully considered (Berman 2021; Clarke 2023; Ward et al. 2013). Other challenges include its complexity and the lack of detailed information on the data collection process, impacting accessibility for researchers due to the unstructured nature of the data (Hopp et al. 2019).

NLP tools have not been widely applied in PEA to date. Recent developments in LLMs could potentially contribute to the field's advancement. This is particularly relevant for languages other than English, as the field has been dominated by English-language data, limiting the inclusion of diverse sources. Earlier works identified challenges in employing NLP for event filtering and classification, such as poor performance and data ambiguities leading to misclassifications and false positives (Hanna 2017; Makarov et al. 2015).

Automated tools also introduce limitations to PEA, such as overreporting due to duplicate entries (Lorenzini et al. 2022). Wiedemann et al. (2022), in their NLP-based framework for German news article detection, emphasize the importance of extensive preprocessing for automatic news identification. They highlight that out-of-sample

recognition of relevant events depends on the training sample, while the length of news content restricts the quality of their model output.

### *Challenges of protest data in authoritarian regimes*

Research on protest data in authoritarian regimes often relies on English-language sources. While translation by international outlets can increase coverage, these agencies may prioritize news aligned with their own interests, often neglecting certain topics or regions (Dollbaum 2021). Moreover, journalists in restrictive environments face state censorship and other risks, leading to self-censorship that suppresses reporting on sensitive topics such as protests (Chang and Manion 2021; Dollbaum 2021; Ong 2021).

To address these limitations, data-gathering initiatives in authoritarian states have used various sources, including social media (Zhang and Pan 2019), activist networks (Lankina and Tertytchnaya 2019), news reports (Bizyukov and Dollbaum 2021; Salehyan et al. 2012), and multi-source methods combining reports and case studies (Raleigh et al. 2010; Sundberg and Melander 2013). As a result, existing datasets tend to focus on specific timeframes or prioritize particular event categories, such as labor protests (Bizyukov and Dollbaum 2021) or violent clashes (Sundberg and Melander 2013), depending on the project's research goals.

## Creating the Russian contentious events dataset (RCED)

The 2010s and 2020s represent a crucial period in Russian history, as the political regime systematically implemented restrictive legislation, adopted new methods of repression, and exerted increasing control over political expression (Lewis 2020; Mamaev 2024). Understanding the evolution of the country's contentious politics amid this growing authoritarianism – particularly during Vladimir Putin's third and subsequent presidential terms – requires reliable and granular data, yet existing datasets offer conflicting and fragmented accounts. State censorship, media control, and the constant risk of repression for activists and journalists make it difficult to obtain data that accurately capture changes in contentious action. To overcome these limitations, this research introduces a novel methodology for analyzing user-generated data from Twitter (X). This paper demonstrates how leveraging social media data and LLMs can automate the data collection process, creating a more systematic, replicable, and less resource-intensive alternative to traditional approaches.

### *Choosing the platform for data collection*

Russian journalists face assault, intimidation, persecution, and even murder, leading to self-censorship and limited coverage of sensitive topics such as protests (Bodrunova, Litvinenko, and Nigmatullina 2021). State censorship and control over traditional media further restrict the flow of information, while media blackouts intentionally suppress coverage of certain events. The existing opposition media have faced numerous threats and challenges. They have been classified as "foreign agents" and subjected to sanctions that hinder their ability to operate within Russia and access essential resources (Goncharenko and Khadaroo 2020).

In contrast to traditional media, social media platforms offer a space for more open discussion and information-sharing. While the Russian regime has attempted to control platforms such as VK (Pan 2017), it has been less successful in censoring alternatives such as YouTube, Facebook, and Twitter. These platforms enable bloggers, opposition media, and individual activists to report on issues neglected by state-controlled outlets, publish independent information, and build communities critical of the regime (Alieva, Moffitt, and Carley 2022; Denisova 2017). Therefore, this study uses social media as its primary source for PEA data collection, leveraging its potential to capture a broader range of events and perspectives amid growing restrictions on traditional media and civil society.

Modern traditional media organizations maintain a social media presence, and individual users publish relevant information. However, data collection from these platforms faces significant restrictions. Access is typically governed by application programming interfaces (APIs), but these tools often limit data availability, functionality, and usage (Littman et al. 2018). For example, following the Cambridge Analytica scandal, Facebook (Meta) heavily revised its data access policies for researchers (Venturini and Rogers 2019). Similarly, Twitter (X) retained API access but introduced fees that are often prohibitive for many academics and developers (Chang et al. 2023). These limitations frequently lead researchers to resort to scraping, a practice that raises its own technical and ethical concerns (Trezza 2023).

Selecting a suitable platform for data collection requires considering internet penetration rates, network prevalence within a country, and user demographics (Weidmann and Rød 2019a). A network's prevalence influences its reach, which can affect reporting bias and the type of data collected. For example, Facebook and Twitter are more likely to contain posts on sensitive topics because these platforms face less control from the Russian regime. Conversely, protest-related posts are less common on VK, a state-censored platform associated with the highest number of prosecutions for online activity (OVD-Info 2024).

The choice of data source also depends on research objectives and the desired data format. Since this research aims to generate a dataset of reports on contentious action, platforms focused on textual and image-based content are more suitable than those centered on video, such as YouTube (Zhang and Pan 2019). A platform that prioritizes text and images is therefore a more appropriate choice for identifying and analyzing reports of contentious events.

Finally, creating a dataset spanning a long timeframe requires access to historical data. The chosen platform must therefore have a sustained presence, a stable user base, and accessibility within the country – even if, as in Russia, access requires a VPN (Ramesh et al. 2020). Practical data accessibility is another critical factor. Facebook, for instance, presents significant challenges due to its complex API access procedures and data export limitations. Telegram, while more accessible, did not cover the full timeframe required for this study.

Twitter (X) emerged as the most suitable option, offering straightforward data access via its API and third-party tools such as the Python wrapper Twarc2.[2] Despite its lower usage compared to VK, X remains a vital platform for opposition voices, political discourse, and international audiences (Alieva, Moffitt, and Carley 2022; Dollbaum 2021). Its regional user clusters and communities foster consistent engagement with politics and

contentious events (Alexanyan et al. 2012), making it a medium for "discursive struggles" between state and opposition forces (Dehghan and Glazunova 2021, 743). X's network includes media organizations, individual activists, and eyewitnesses, providing access to valuable data on events that are often underreported by traditional media due to their remote location or intentional blackouts.

X's format enables users to post concise messages containing specific information. In addition to the tweet's text, user details, and timestamp, the platform provides data on user interactions such as likes, reposts, quotes, and replies. This interaction data allows for more elaborate methods, such as network analysis, to identify clusters of users engaging with specific posts and topics (Riquelme and González-Cantergiani 2016).

## *Initial data collection*

The analysis aimed to identify reports of protest events on specific dates, posted by media outlets, eyewitnesses, or participants. The Twitter search used the keywords "protest" and "rally" (*protest* and *miting* in Russian) to collect posts referring to such events. While other keywords were considered (e.g. *progulka*, a term for peaceful marches), they largely yielded irrelevant results or duplicates of posts already captured by "protest" or "rally," justifying the final selection.[3] This search resulted in 5,713,892 tweets and retweets from 1 January 2010, to 31 March 2023. Retweets were retained to facilitate network analysis, which helps identify information dissemination patterns and potential user biases.

The keyword "protest" often appeared in tweets discussing opinions or arguments rather than reporting on actual events. This noise, combined with the sheer volume of data, numerous duplicates, overreporting of major incidents, and the inclusion of protests outside Russia, made manual filtering impractical. To address these challenges, this study employed computational techniques, including keyword analysis, topic modeling, and duplicate removal, to isolate relevant protest events within Russia.

## *Selecting the model, fine-tuning, and deployment*

An ML approach using existing NLP models was the most suitable method for this task. This method involved a binary classification of tweets based on relevance. A tweet was classified as relevant if it reported on a specific protest, providing its location (city and a more precise venue) and, ideally, other details such as the date, organizers, attendance,

**Table 1.** Sample tweets from the dataset.

| Date (y/m/d) | Translated text | Class | Reason |
|---|---|---|---|
| 2010–06-05 | Tajik students staged a rally in front of the Uzbek Embassy in the United States | 0 | Protests in the United States |
| 2010–03-06 | People in the office who wanted to attend a rally for reform of the Ministry of Internal Affairs cut off the phone. Would like to know what time it starts. | 0 | No location |
| 2010–10-24 | By 13:00 I am going to Labor (Freedom) Square for a rally in support of Yegor Bychkov #ekb | 1 | Location and topic |
| 2010–11-09 | Rally of paratroopers on November 7, 2010 | 0 | No location |
| 2010–09-29 | Strange rally in Serpukhov | 1 | Location in Russia |

**Table 2.** Performance metrics of the fine-tuned RuBERT model for binary contentious event classification.

| Class | Precision | Recall | F1 score | Support |
|---|---|---|---|---|
| 0 | 0.92 | 0.91 | 0.91 | 1,412 |
| 1 | 0.79 | 0.80 | 0.80 | 588 |
| Accuracy | | | 0.88 | |
| Macro avg. | 0.85 | 0.86 | 0.85 | 2,000 |
| Weighted avg. | 0.88 | 0.88 | 0.88 | 2,000 |

*Notes*: Fine-tuning was conducted over 61 epochs with a batch size of 64, a learning rate of 1e-03, and AdamW as the optimizer. Tweets were assigned to classes based on a confidence interval of 0.4, which was determined to be the optimal threshold for balancing precision and recall based on performance scores and manual evaluation of the model's output.

or issues raised. Tweets were classified as irrelevant if they lacked specific locational details, contained only ambiguous content (e.g. single words or emojis), or discussed an unrelated topic. To illustrate this classification, Table 1 presents translated examples of relevant (1) and irrelevant (0) tweets.

To conduct an automated binary classification of tweets, the Bidirectional Encoder Representations Transformers (BERT) model developed by Google was chosen (Devlin et al. 2018). This state-of-the-art open-source model allows researchers to train or fine-tune its parameters for specific tasks and data types. BERT performs well with shorter texts and text classification tasks (González-Carvajal and Garrido-Merchán 2023). Since the tweets collected for this project were in Russian, this research employed a pre-trained Russian-language initialization of BERT – RuBERT. Trained on Russian-language data, RuBERT enhances performance and increases the accuracy of results when working with Russian texts (Kuratov and Arkhipov 2019).

To fine-tune the classification model, a gold-standard dataset was manually created by a single coder based on the location criteria. To prevent overfitting, this dataset was incrementally expanded to a final size of 6000 training and 4000 validation tweets, which were randomly selected. The model's performance was evaluated using precision, recall, and the F1 score, with detailed metrics presented in Table 2. The model successfully classified both relevant (F1 = 0.80) and irrelevant (F1 = 0.91) tweets. A qualitative assessment was also conducted to refine the confidence threshold for classification. Finally, the fine-tuned model was applied to the entire dataset (excluding retweets) to identify all relevant posts.

### *Geographic entity recognition*

The tweets classified as relevant were further processed to extract geographic entities and pinpoint the locations of the identified contentious events. For this task, Geographic Entity Recognition (GER), GPT-3.5 was chosen due to its extensive knowledge of geographic locations, street names, and landmarks, acquired through training on vast amounts of data. Its ability to handle misspellings common in social media data and its contextual understanding of idiomatic and colloquial expressions further enhance its suitability for GER (Yin, Li, and Goldberg 2023; Zhang et al. 2024). This is beneficial for identifying locations in tweets where names are often shortened or presented as hashtags.

The model first extracted names of cities and specific locations (e.g. squares, shopping malls) from the tweets. These entities were then manually reviewed to correct errors and remove false positives – such as locations that could not be geographically identified – that were misclassified by the RuBERT model. Next, Google's Geolocation API was used to assign each verified location to its corresponding Russian administrative region and obtain coordinates for mapping. The API query excluded Crimea and other internationally unrecognized territories claimed by Russia after its 2022 full-scale invasion. Any remaining overseas territories were filtered out through a combination of dictionary look-ups and manual review of the region names provided by the API.

### *Duplicate removal*

To avoid overreporting of protests, particularly large-scale events reported by multiple media sources and individuals, duplicate entries were removed based solely on their textual content. Emojis and punctuation were excluded from this comparison, as some posts were identical in wording, but differed in their use of emojis or additional punctuation, such as exclamation marks or capitalization.

Removing contextual duplicates – tweets that were semantically similar but not textually identical – was a more complex challenge. To address this, RuBERT embeddings were used to perform weekly pairwise comparisons of all tweets. Any tweets with a semantic cosine similarity above 80% were classified as duplicates and removed (Isotani et al. 2021; Reimers and Gurevych 2019). In a second step, reports of the same event were merged. Tweets sharing the same location and occurring within a 24-hour window were consolidated into a single entry. This approach was based on the assumption that multiple major events are unlikely to occur in the same locality simultaneously. Merging these reports not only reduced redundancy but also created a more comprehensive record of each event by integrating details from multiple sources.

While this approach risks underreporting simultaneous events in the same location, such occurrences are rare outside major cities like Moscow and St. Petersburg. To mitigate this risk further, a contextual check was performed during the summarization stage. If multiple distinct locations or topics were detected within a single group of aggregated tweets, the model was prompted to generate separate summaries for each, which were then divided into distinct dataset entries.

## RCED and contention in Russia

The final dataset comprises 59,081 reports on contentious events, each containing the date, locational details (region, city, and specific place), and geocoordinates. Protecting user anonymity was a critical step. While tweet and user IDs were removed, this measure alone is insufficient, as original text can be used for re-identification (Sousa and Kern 2023). Therefore, a two-step process was applied to generate non-identifiable event descriptions: the original Russian tweets were first translated using the Google Translation API and then summarized using GPT-3.5. This method produced a fully anonymized, English-language dataset of event summaries, with no original Russian

content retained. The resulting dataset is well-suited for multiple applications: the location data facilitate regional analysis and mapping, while the anonymized descriptions enable event classification and qualitative assessment.

### *Key reporting actors*

The retweets obtained during the initial data collection allowed for the identification of various user categories that reported on protests using Twitter (X). Knowing the categories of contributors also facilitated the potential identification of biases in event descriptions and provided a deeper understanding of the sources. Figure 1 depicts the network of the most influential users whose posts contributed to the dataset. This network, constructed using retweets, quotes, and replies as interaction metrics, highlights which users' tweets were more influential in disseminating information about protests. As established by prior research, the extent of user engagement with content serves as an indicator of their influence within a social network (Cha et al. 2010; Riquelme and González-Cantergiani 2016).

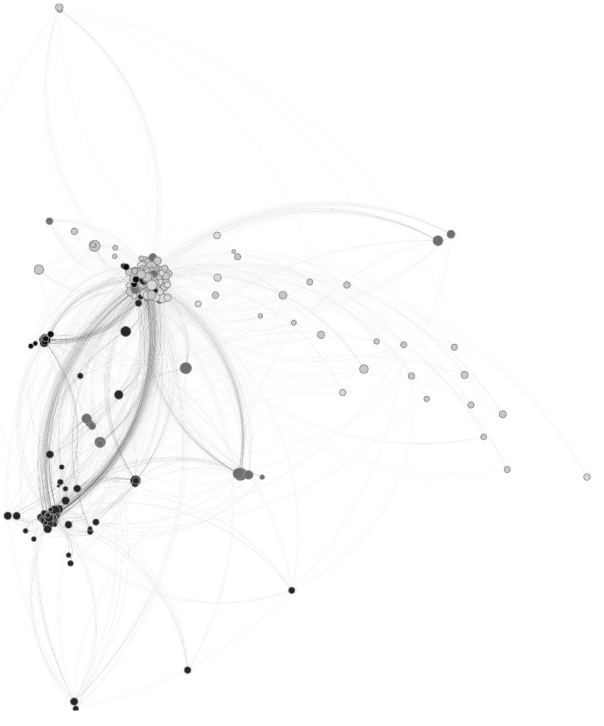

**Figure 1.** Network of influential users contributing to the dataset. *Notes*: the central light-grey cluster (●) comprises anti-regime opposition media, journalists, and activists. The dense dark-grey cluster (●) in the bottom left represents pro-regime media, while the black cluster (●) closer to the center-left consists of Communist party deputies and supporters. The grey cluster (●) on the bottom right represents opposition bloggers. The remaining notable users include deputies from opposition parties (right-hand side), pro-regime supporters and activists (bottom-left and center), and Ukrainian media (top center-left).

While this figure does not encompass the entire network of users contributing to RCED, it shows the core representative categories of users who generated reports on contentious events. To create this visualization, the full network of influential users was first manually examined, and the accounts were classified into distinct categories based on a qualitative assessment. The network was then simplified into the directed bipartite graph shown, which comprises 352 visible nodes and 6906 edges. The node clusters represent influential users, with node size corresponding to the number of interactions an account received. Colors are assigned based on modularity, a measure that detects well-defined communities within the network. The edges illustrate interactions (retweets, quotes, and replies) between influential users and the accounts that engaged with them. The proximity of clusters indicates frequent interaction between those user categories.

Significant bot activity has been previously documented on Russian Twitter (Alieva, Moffitt, and Carley 2022), and this study's methodology mitigates its impact. The initial RuBERT classification filters for tweets containing factual information about an event, effectively removing much of the narrative-driven content typical of bot networks. The duplicate removal and summarization steps further focus on objective data, excluding subjective discussions. The final summaries retain only key details (e.g. date, location, topic, and actors) and minimize the influence of any remaining bot-generated discourse on final dataset entries.

The graph suggests that opposition activists, journalists, and media played a significant role in generating reports on contentious events for the dataset. However, the presence of pro-regime activists, Communist Party supporters, politicians, and other actors shows the diverse range of users contributing to protest-related discussions.

### *Event frequency and classification*

The wide range of events in RCED allows for an analysis of how contentious action evolved. The descriptions of protest reports enable a better understanding of these changes, both longitudinally and at the event level. To demonstrate this, events were classified into three main categories: anti-regime (opposing state policies or challenging authorities); pro-regime (supporting the government and its decisions); and non-political (associated with issues like construction, development, or labor). This classification is based strictly on the explicit information within the event summaries; consequently, it may not capture underlying political dimensions when the source text provides insufficient context for such nuances.

Events were placed in an "unknown" category if the model could not classify them, typically due to ambiguous descriptions or topics falling outside the provided definitions. Beyond this classification, the dataset also includes additional factual information where identified, such as protester numbers and mentioned organizations. All attributes and classification definitions are detailed in the associated codebook.

Figure 2 shows a notable increase in anti-regime events, with spikes corresponding to key periods of contention in Russia. The spikes are particularly evident during the 2011–2012 State Duma elections; the anti-corruption and pension reform protests of 2017–2018; and the protests surrounding the 2019 Moscow Duma elections and Alexei Navalny's poisoning and return in 2020–2021. The graph also shows a high volume of pro-

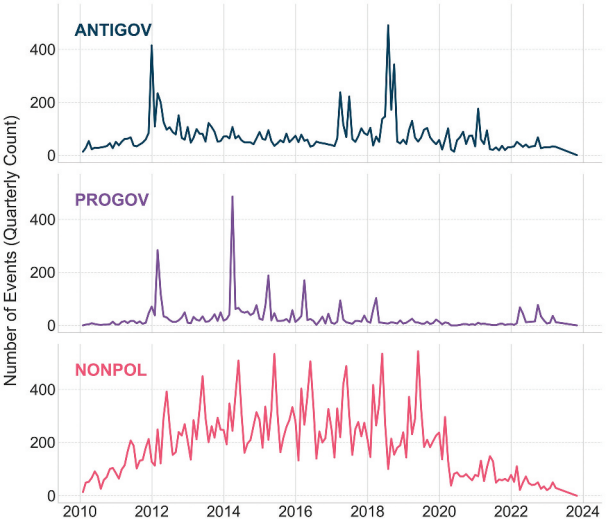


**Figure 2.** Line chart representing monthly changes in the number of RCED-reported events of anti-, pro-regime contention and non-political protests over the 2010–2023 period in RCED.

government rallies, especially during the 2012 presidential election and the 2014 annexation of Crimea, followed by a gradual decline in such reports.

In contrast to political rallies, non-political events exhibit a strong seasonal pattern, capturing regularly occurring celebrations and commemorations such as Labor Day and Victory Day marches, Day of Remembrance and Sorrow, and Day of Memory of the Victims of Political Repression, among others. While the number of these events remained relatively stable from 2014 to 2020, they dropped in parallel with anti-regime events thereafter. This is part of a broader decline in contention across all categories from 2020 onwards, a reduction possibly associated with COVID-19 lockdowns and the subsequent crackdown on activists and social movements. The dramatic post-2020 fall in events, including those concerning labor, social, and economic grievances, indicates that the overall tolerance for mass gatherings of any kind has decreased.

Examining specific periods reveals more granular information. For instance, despite an increase in pro-regime events supporting the annexation of Crimea, the data also show

**Table 3.** Sample of anti-war protest event data in March 2014.

| Date (y, m, d) | City | Region | Category | Subclass | Description |
|---|---|---|---|---|---|
| 2014–03-03 | Uglich | Yaroslavl′ Oblast | ANTIGOV | WAR | Tomorrow at 16:00, a rally will be held on Assumption Square in Uglich in support of the people of Ukraine. |
| 2014–03-11 | Angarsk | Irkutsk Oblast | ANTIGOV | WAR | A rally in support of Ukraine was held in Angarsk. |
| 2014–03-12 | Serpukhov | Moscow Oblast | ANTIGOV | WAR | A rally in support of the people of Ukraine and the Autonomy of Crimea was held in Serpukhov. |
| 2014–03-12 | Ussuriysk | Primorskii Kray | ANTIGOV | WAR | On 12 March 2014, a rally in support of the people of Ukraine took place. It was held from 17:00 to 18:00 on the central square of Ussuriysk. |
| 2014–03-13 | Pervouralsk | Sverdlovsk Oblast | ANTIGOV | WAR | A rally in support of the people of Ukraine was held today in Pervouralsk. |

anti-invasion rallies in cities such as Murmansk, Samara, and Voronezh, and in smaller towns such as Angarsk, Ussuriysk, Uglich, Serpukhov, and Pervouralsk. RCED often includes the exact
time and location for these events (Table 3).

### *Topic distribution*

The majority of anti-regime events addressed the topic of civil liberties, particularly protests against restrictions on freedom such as censorship (e.g. events against the blocking of Telegram). Local politics was also a dominant category, alongside events demanding reforms, policy changes, or opposition to such changes (e.g. opposition to the Yarovaya Law). While some categories may intersect – for example, many protests classified under civil liberties also called for the release of political prisoners and were directed at the president – this classification reflects the most dominant themes in anti-regime protests across the timeframe (Figure 3).

Most pro-regime events were classified as nationalist, i.e. evoking national pride and supporting Russia as a nation, particularly during the annexation of Crimea, the imposition of sanctions, or other international events. The foreign policy category included protests outside embassies, such as demonstrations against Japan concerning the Kuril Islands. A large portion of pro-regime events also took place in support of President Putin or the "reunification" of nations, as also shown in Figure 3. Non-political events were often commemorative (e.g. May 1 and May 9 marches), but also included a large share of environmental protests, housing disputes (concerning poor conditions or defrauded shareholders), and protests over infrastructure (such as the construction of a new church in central Yekaterinburg).

The frequency chart in Figure 4 shows that the prevalence of different anti-regime topics varied over time. While events supporting civil liberties, opposing the government, or concerning local issues were relatively evenly distributed, anti-corruption protests, in contrast, peaked in 2017 and remained at low levels in other years, often being absorbed into broader anti-government themes. A similar trend is visible in the reform category, where the 2018 pension reform protests dominate a category that otherwise saw little activity. This suggests that while some anti-regime rallies are tied to specific issues (such as the Bolotnaya prisoners) or election periods, civil liberties and general anti-government sentiment are more consistent themes.

Non-political events, on the other hand, showed more consistency, with most topics peaking in 2014–2018 before declining (Figure 5). Pro-government events peaked in 2012–2014 and then declined across all categories, potentially suggesting that the regime did not prioritize sustained mobilization, instead organizing rallies only for specific occasions (Figure 6). Further details on each category and the classification process are provided in the codebook.

### *Regional distribution*

The dataset reveals that the majority of anti-regime events took place in Moscow, St. Petersburg, and Sverdlovsk Oblast, as well as Samara Oblast, Khabarovsk Krai, and Moscow Oblast, as illustrated by the treemaps in Figure 7. Pro-regime events were also

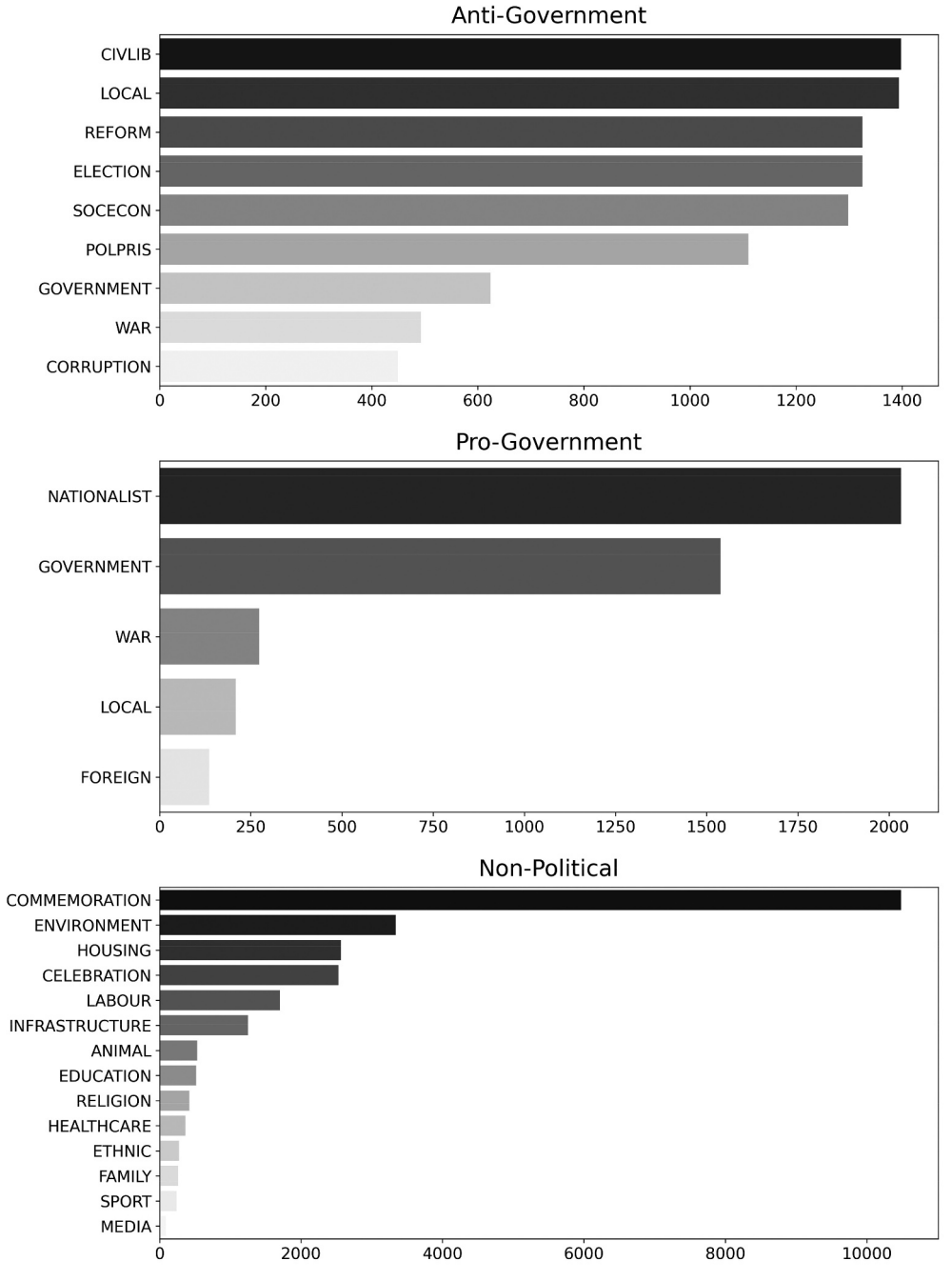


**Figure 3.** Topics prevalent in different categories of events across the RCED dataset.

concentrated in Moscow, St. Petersburg, and Moscow Oblast, with Sverdlovsk Oblast, Krasnodar Krai, and Rostov Oblast also recording high numbers of such events. Non-political events prevailed in Moscow and Moscow Oblast across the entire timeframe, followed by St. Petersburg, Krasnodar Krai, and Sverdlovsk Oblast. Overall, RCED recorded

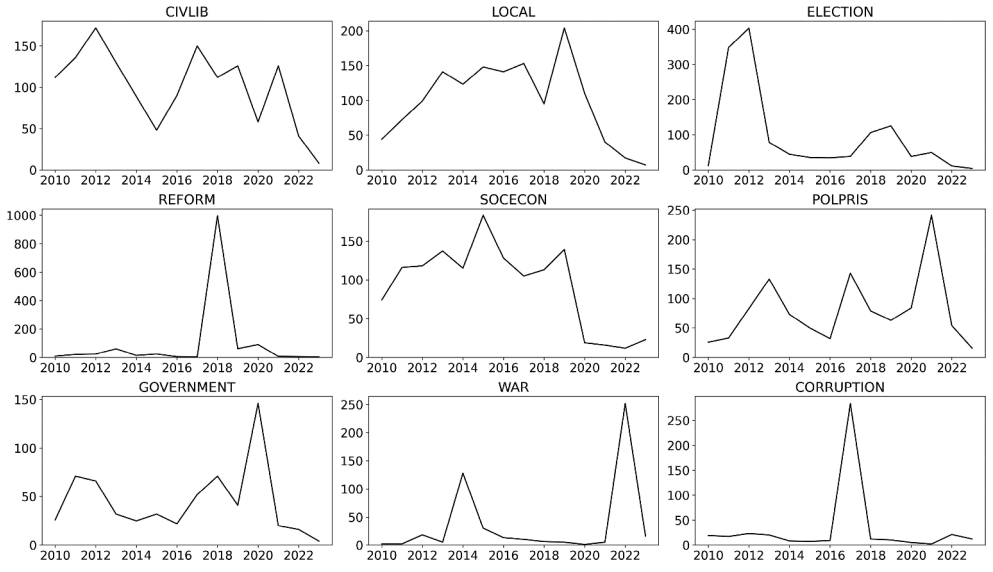


**Figure 4.** Frequency of protests in each identified ANTIGOV topic.

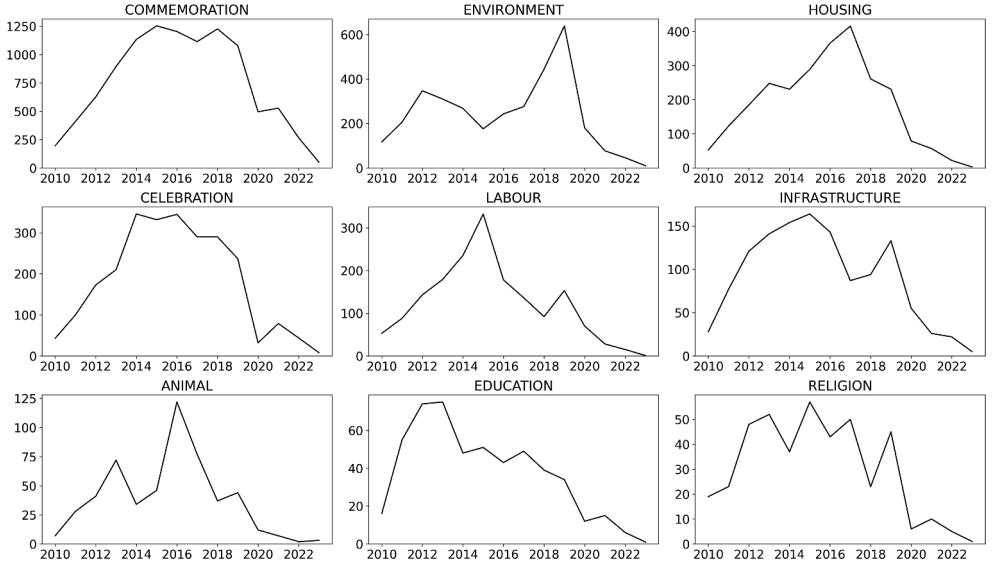


**Figure 5.** The frequency of protests in each identified NONPOL topic.

more anti-regime events than pro-regime demonstrations. Notably, the regions with the highest incidence of anti-regime contention also tended to dominate in non-political events concerning economic grievances, labor, or local policies.

To explore similarities in protest dynamics across Russia's regions, this study employed Uniform Manifold Approximation and Projection (UMAP). For this method, each region was represented by its time series of monthly protest counts, forming

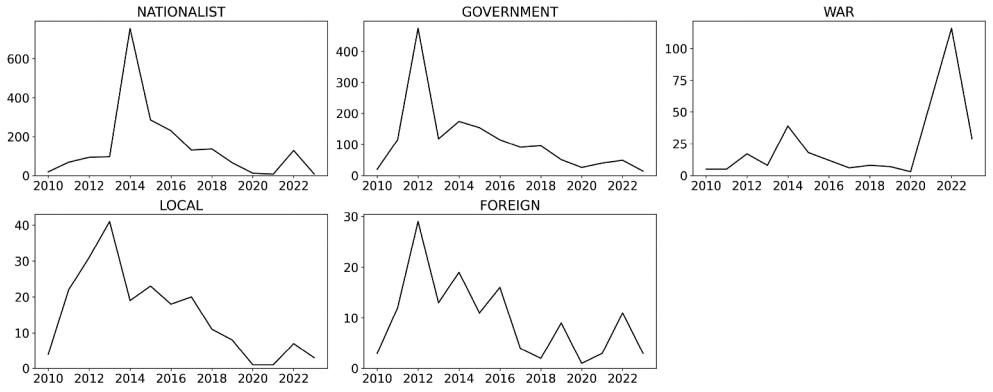


**Figure 6.** Frequency of protests in each identified PROGOV topic.

a 160-dimensional vector. UMAP was chosen for its advanced ability to visualize complex, non-linear data and improve clustering (McInnes et al. 2018; Pealat, Bouleux, and Cheutet 2021). By projecting these high-dimensional data into a two-dimensional space, the method reveals distinct groups of regions that share similar protest patterns, as shown in Figure 8.

The analysis identified six clusters, which are described in the online Appendix. It revealed a high similarity among the most populous and economically developed regions, including Moscow, St. Petersburg, Sverdlovsk Oblast, Novosibirsk Oblast, Samara Oblast, and Tatarstan, among others (the "Contentious Regions" cluster, located in the bottom left of the plot). This widespread similarity is driven by nationwide protest waves, such as those organized by Navalny's Anti-Corruption Foundation (FBK), demonstrations against pension reform, and rallies against federal election manipulation.

At the other end of the spectrum are regions of "contained contention" – those with the lowest recorded number of events, such as Chechnya, Ingushetia, Chukotka, and Kamchatka. These regions clustered with other areas that are mostly remote and sparsely populated, including Magadan Oblast, the Mari-El Republic, and Amur Oblast. They exhibit the lowest protest incidence in the dataset, showing little change even during nationwide events. This pattern may reflect several factors, including the region's remoteness, its perceived political significance, and the state of local activist networks, such as the presence of FBK representatives or active party offices (e.g. Yabloko, CPRF).

Between these two extremes lie several other clusters. These include regions with consistent but moderate contention (e.g. Omsk, Stavropol'), "special-case" regions (the conflicted Republic of Dagestan and the enclave of Kaliningrad Oblast), and "event-driven" regions (e.g. Tomsk, Buryatia). While these middle-ground regions participated in some nationwide protests, their contentious activity was generally lower, more focused on local issues, and covered a limited range of topics.


**Figure 7.** Regional incidence of protest events within each RCED category.

## *Comparing with existing datasets*

### *Dataset description*

For comparison, RCED is benchmarked against several key datasets covering a similar period. The first is LaRUPED, a manually coded dataset of 5824 events (2007–2016) derived from an opposition website, which is positioned as a supplementary source rather than a complete record (Lankina and Tertytchnaya 2019). Two large-scale, automated datasets, GDELT and ICEWS, are also included. GDELT continuously catalogs

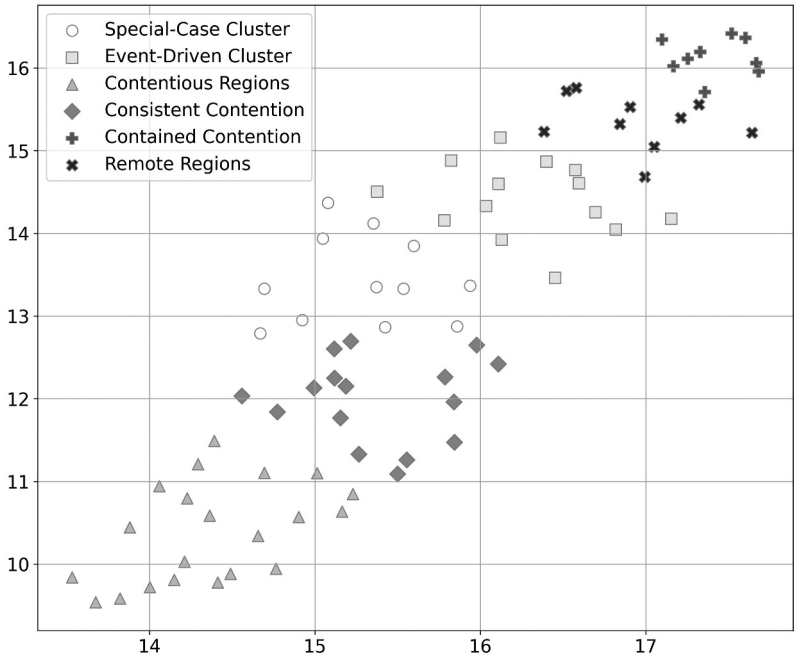


**Figure 8.** UMAP built based on calculated differences in contentious event trends across all Russian regions.

events from global news but counts articles rather than unique events, leading to documented overreporting and duplication issues (Althaus, Peyton, and Shalmon 2022; Clarke 2023). ICEWS, though discontinued in 2023, remains widely used but shares similar limitations (Boschee et al. 2015; Lorenzini et al. 2022). POLECAT, the successor to ICEWS as of 2023, is a machine-coded dataset using modern NLP models on international news reports, with data available from 2018 onwards (Halterman et al. 2023; Scarborough et al. 2023). Finally, the Mass Mobilization Protests Dataset (MMAD) employs a hybrid approach, using machine learning to pre-filter reports from three news agencies, which are then manually coded by experts (Weidmann and Rød 2019a). Its primary documented limitation is underreporting (Clarke 2023; Dollbaum 2021).

### *Event frequency*

Figure 9 presents the event count comparison among the datasets. RCED, GDELT and ICEWS, cover the 2010–2023 period, MMAD covers the 2010–2020 timeframe, while LaRUPED and PLOVER cover 2010–2017 and 2018–2023, respectively.

The datasets show some consensus, particularly in capturing the major protest wave during the 2011–2012 election cycle (Koesel and Bunce 2012). However, significant divergences appear around the 2014 annexation of Crimea. GDELT reports a massive surge in pro-regime rallies (Hale 2018), a spike not observed in ICEWS or LaRUPED. Following this, RCED, GDELT, and MMAD show sustained reporting from 2014 to 2017, while ICEWS shows a decline. GDELT also uniquely emphasizes the 2022 full-scale invasion protests, though

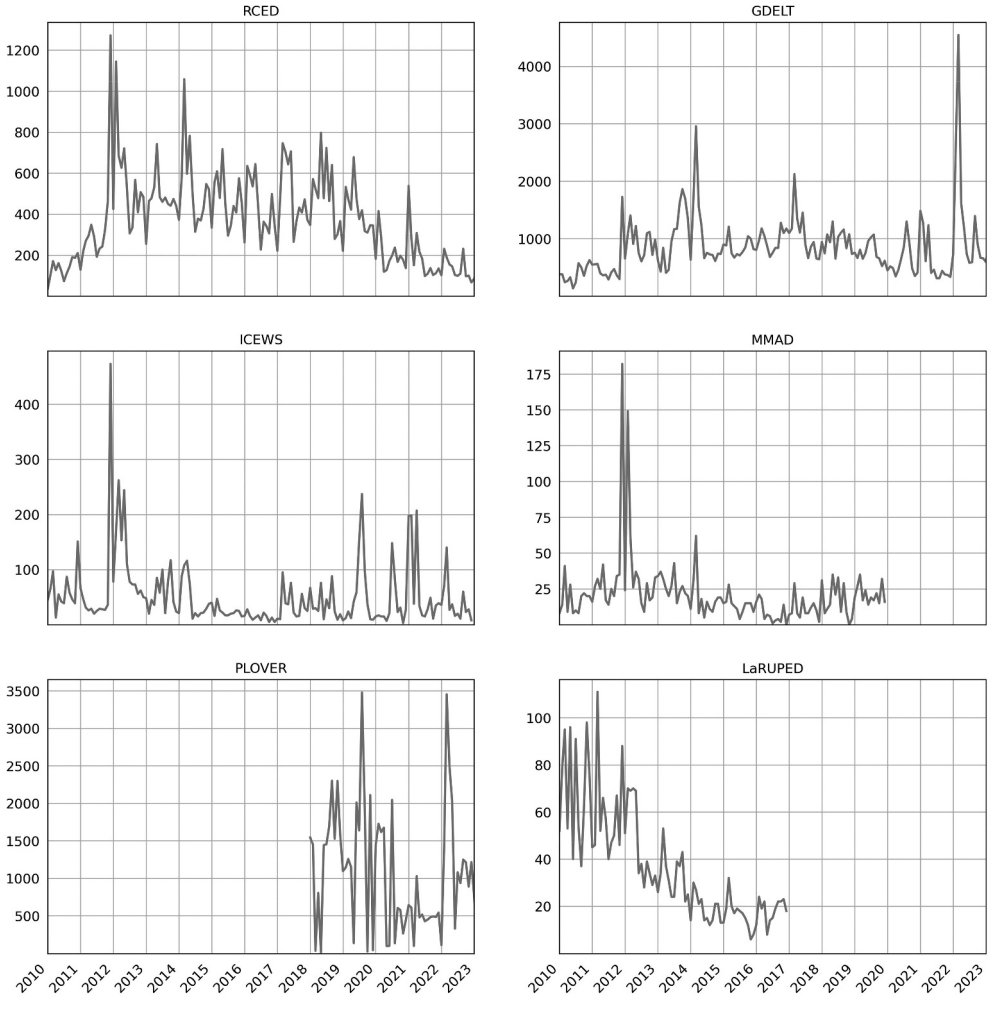


**Figure 9.** The comparison of the datasets on protests in Russia that encompass a wide range of events and at least partially cover the 2010–2023 timeframe.

their actual scale was likely suppressed by intense state repression (Dubina and Arkhipova 2023).

These event-specific differences reflect broader disagreements in long-term trends. RCED, alongside the manually coded LaRUPED and MMAD, indicates a general decline in contention after the 2011–2012 peak. In contrast, the automated GDELT and ICEWS datasets suggest protest frequency either increased or remained stable, with ICEWS and POLECAT showing more activity in later years.

### *Trend comparison*

To quantitatively compare protest trends, Euclidean distance was calculated on the monthly time-series data from each dataset. This metric was chosen for its suitability in time-series analysis and its ability to directly measure the absolute difference between trends at corresponding points in time (Deza and Deza 2009; He, Agard, and Trépanier

**Table 4.** Pairwise Euclidean distances between datasets.

| | RCED | GDELT | MMAD | ICEWS | LaRUPED | PLOVER |
|---|---|---|---|---|---|---|
| RCED | 0 | 0.051354 | 0.079553 | 0.084899 | 0.097584 | 0.113631 |
| GDELT | 0.051354 | 0 | 0.094246 | 0.086381 | 0.103002 | 0.097953 |
| MMAD | 0.079553 | 0.094246 | 0 | 0.077058 | 0.110099 | 0.171710 |
| ICEWS | 0.084899 | 0.086381 | 0.077058 | 0 | 0.115484 | 0.146421 |
| LaRUPED | 0.097584 | 0.103002 | 0.110099 | 0.115484 | 0 | NA |
| PLOVER | 0.113631 | 0.097953 | 0.171710 | 0.146421 | NA | 0 |

2020). To ensure a fair comparison between datasets with vastly different total event counts, the monthly protest counts for each were first normalized into proportions of their respective totals. The distance was then calculated as the square root of the sum of squared differences between these normalized monthly proportions, performed only on the overlapping period for each pair of datasets. The Euclidean function from SciPy was used for this analysis, with the results presented in Table 4.

GDELT, MMAD, and ICEWS demonstrate the highest similarity to RCED's monthly event trends, while LaRUPED and PLOVER exhibit the greatest distance. Although GDELT's overall trend differs from RCED's, the magnitude of its monthly fluctuations aligns closely, particularly in highlighting periods of heightened contention. This pattern of similar fluctuation magnitudes – despite variations in absolute event counts – is also reflected in the MMAD and ICEWS datasets. Conversely, PLOVER shows a higher level of disagreement, particularly with MMAD; however, this discrepancy may be attributed to the limited two-year overlap between these datasets. The "NA" score between PLOVER and LaRUPED indicates no overlapping years.

Overall, the datasets are different in scope, reflecting different trends over longer timeframes and in the dynamics of change over time. These variations can potentially impact the quality of analysis and may have severe theoretical and empirical implications for understanding the evolution of contentious action. Unfortunately, the accuracy of each dataset cannot be definitively determined through comparison alone due to the absence of an objective, accurate account of the events that took place, which is one of the biggest challenges in PEA data.

## *Limitations*

Even though RCED aims to provide an accurate estimate of contentious events based on the existing reports, some critical issues should be addressed in further research and datasets. One of the most significant problems is overreporting, where specific instances of contentious action may be mentioned multiple times. Although this dataset employs a technique for duplicate removal, contextual removal is not always feasible due to varying descriptions and the concise nature of tweets, often making it difficult to distinguish between separate events and duplicate reports. While techniques like duplicate detection based on matching and contextual embeddings are employed, they may not fully address the issue.

Underreporting of specific events is another key limitation, potentially stemming from misclassifications by the RuBERT and GER models. Furthermore, on a more fundamental level, RCED inherits the biases of its source platform. X's popularity is concentrated in urban areas, leading to a likely underrepresentation of events in smaller regions. This

focus on a single platform means the dataset cannot account for events reported exclusively on other services, such as Telegram and WhatsApp groups, which may be more popular in certain regional or demographic contexts in Russia.

Another important consideration is the changing nature of user preferences and the popularity of social media platforms. As Fiesler and Dym (2020) point out, online communities and platforms experience periods of popularity and mass exodus due to platform-based reasons, changes in design and policies, and shifting communication preferences. Moreover, state decisions, such as the blocking of Twitter in 2022, can impact a platform's individual and regional use. While this may not necessarily affect dedicated opposition communities, such changes in usage patterns can change event reporting. This variability must be considered when using social media as a source for PEA data, particularly over longer timeframes.

While many contentious events are reported on the actual date of occurrence, some tweets mention events before they happen (as announcements or expressions of intent to attend) or after they have already taken place. Since the dataset considers the date of the report to be the date of its occurrence, the accuracy of daily event counts may be affected, leading to potential over- and underreporting and inconsistencies, particularly on a daily timeframe. Therefore, the exact dates of specific events should be verified through other sources whenever possible, especially for major events.

Although RuBERT, GPT-3.5, GPT-4, and Google Translate are state-of-the-art tools that can provide reliable results, the lack of systematic literature on their application in PEA requires careful attention and manual verification to address potential inconsistencies in the dataset. While the use of more recent and sophisticated models may improve the quality of such work, issues such as false outputs, factual inaccuracies, hallucinations, and failure to adhere to instructions can be difficult to detect in larger datasets, regardless of the model (Augenstein et al. 2024). Opaque training practices also limit reproducibility and bias mitigation (Sapkota, Raza, and Karkee 2025), which in turn complicates analysis even further.

It is also crucial to consider the compliance of some LLMs with the regulatory frameworks of specific governments and institutions. There are significant concerns regarding data security on proprietary platforms, as their lack of transparency in training practices and data use may prevent researchers in certain jurisdictions from employing them. While it is possible to replace commercial LLMs and Google services with open-source alternatives approved in a particular region, the performance of these models must be carefully assessed. Relying on different models may also raise quality concerns and reduce accessibility, as they often require substantial computational resources, financial investment, and advanced skills for model training.

## Conclusion

This paper presents a comprehensive dataset of contentious events in Russia from 2010 to March 2023, automatically generated using existing computational tools. Analysis of this dataset offers valuable insights into the evolution of contentious action in Russia during a period of declining freedoms and rising authoritarianism. It reveals that protest

dynamics were largely event-driven and highly localized, while pro-regime participation coalesced around state-driven events.

The data show a significant reduction in all forms of public contention after 2020. This decline affected all event types, including non-political grievances, demonstrating the broad impact of repression on activists and social movements. The analysis also identifies a core of highly active regions where most national events occurred, alongside a large cluster of outlier regions where contention was virtually absent – a pattern attributable to factors including local governance, geographic remoteness, and weak political networks. RCED demonstrates that modern computational tools offer resource-efficient solutions for data collection and that social media can serve as a valuable source for studying such events.

With its coverage of protest events, RCED enables the identification of protest trends at both federal and regional levels. Descriptions derived from X posts by opposition leaders, politicians, and pro-regime accounts provide a better understanding of how protests evolved in Russia and how contentious action declined over 2010–2023. By dividing the events into three categories, the dataset allowed for event classification based on topics of claim-making using BERTopic, a state-of-the-art topic modeling technique. This approach revealed qualitative changes in contentious action and a decline in the range of issues raised during protests in Russia over time.

RCED can be used for longitudinal quantitative and qualitative studies of protest, social movements, and contentious politics in Russia. The methodological workflow itself is not limited to X and can be adapted by researchers to other social media platforms, such as VK, WhatsApp, Telegram, Meta, and others, to expand upon this work. Even though API access to data has been increasingly limited and commercialized, additional methods of data access can be explored. Solutions and tools that comply with local regulations and ethical frameworks, such as data donations (Boeschoten et al. 2022; Northcott et al. 2025; Ohme et al. 2024) or social media data repositories (Acker and Kreisberg 2020), can be considered. The RCED workflow provides a practical case study for examining the advantages and weaknesses of these tools, addressing methodological issues, and designing new techniques to improve data collection. This can help researchers examine the biases and implications inherent in computational methods, especially when working with social media and other digital sources.

However, there are significant limitations to the application of these tools. RCED may lack precision in certain instances, such as specific dates and the number of unique reports. Potential areas for further research include exploring how different social media platforms can serve as sources of PEA data and examining their biases, which may contribute to event over- or underreporting. Additionally, integrating diverse platforms (e.g. Telegram, Facebook, WhatsApp, and YouTube) into broader, longitudinal data collection would strengthen cross-validation capacities. Such an approach is important for moving closer to a more objective assessment of contentious action in Russia and overall differences in social media.

The efficiency and suitability of various tools for identifying protest events and performing GER should be assessed across different contexts and languages. Further investigation is needed to understand how data collection methods, reporting bias, and data quality vary depending on regime type and the presence of specific social media platforms. Combining traditional data collection methods from newspapers and websites

with social media data could provide more comprehensive descriptions and a deeper understanding of the evolution of contentious action and politics.

Future research should explore the use of distance metrics, such as Euclidean distance and more complex measures, in event analysis and computational approaches to PEA. This could contribute to more rigorous hypothesis testing and analyses of contentious trends, an area currently lacking extensive research. While the tools mentioned in this study, such as RuBERT, BERTopic, and GPT, can be helpful, a range of other available tools and techniques could potentially be even more productive and contribute further to both data collection and analysis of large datasets.

Recent developments in ML, NLP, and computational analysis create more research opportunities and may lead to breakthroughs in the field, potentially addressing current challenges in PEA and data collection through interdisciplinary collaboration. Beyond social media, these tools allow for more comprehensive and resource-efficient PEA when working with traditional sources that often provide richer contextual detail, such as newspapers in both digital and physical versions. Using optical character recognition combined with the improved textual analysis capacity of LLMs to re-examine issues of selection, reporting, and sampling bias could dramatically improve our data collection methods, facilitate the exploration of similar biases in other data sources, and help collect more representative and comprehensive data, ultimately contributing to studies of protest, social movements, and contentious politics.

## Notes

1. https://www.gdeltproject.org/.
2. The dataset was collected before Twitter (X) introduced restrictions on data collection for researchers in May 2023.
3. The use of 'protest' and 'rally' may have become less frequent in 2020–2022 following the rise of solitary pickets amid COVID-19 and other restrictions. However, RCED still captures a wide range of events from this period. While a rapid decline in contention was observed, this trend occurred across all event categories, and there was no evidence that a significant change in protest-related terminology skewed the results. Nevertheless, a longer-term observation (e.g. 2020–2025) might potentially reveal such a shift.

## Author contributions

CRediT: **Bogdan Mamaev:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

Bogdan Mamaev http://orcid.org/0000-0002-1649-6501

## Data availability statement

The dataset, codebook, ANTIGOV, PROGOV, and NONPOL samples are available publicly on Zenodo: Mamaev, B. 2025. "Russian Contentious Event Dataset (RCED)." Zenodo. doi:10.5281/zenodo.17500460.

## References


Acker, A., and A. Kreisberg. 2020. "Social Media Data Archives in an API-Driven World." *Archival Science* 20 (2): 105–123. https://doi.org/10.1007/s10502-019-09325-9

Alexanyan, K., V. Barash, B. Etling, R. Faris, U. Gasser, J. Kelly, J. G. Palfrey, and H. Roberts. 2012. "Exploring Russian Cyberspace: Digitally-Mediated Collective Action and the Networked Public Sphere." *Berkman Center Research Publication*). Accessed December 9, 2025. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2014998

Alieva, I., J. Moffitt, and K. M. Carley. 2022. "How Disinformation Operations Against Russian Opposition Leader Alexei Navalny Influence the International Audience on Twitter." *Social Network Analysis and Mining* 12 (1): 80. https://doi.org/10.1007/s13278-022-00908-6

Althaus, S., B. Peyton, and D. Shalmon. 2022. "A Total Error Approach for Validating Event Data." *American Behavioral Scientist* 66 (5): 603–624. https://doi.org/10.1177/00027642211021635

Andretta, M., and E. Pavan. 2018. "Mapping Protest on the Refugee Crisis: Insights from Online Protest Event Analysis." In *Solidarity Mobilizations in the "Refugee Crisis": Contentious Moves*, edited by D. della Porta, 299–324. Cham: Palgrave Macmillan.

Augenstein, I., T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, et al. 2024. "Factuality Challenges in the Era of Large Language Models and Opportunities for Fact-Checking." *Nature Machine Intelligence* 6 (8): 852–863. https://doi.org/10.1038/s42256-024-00881-z

Benson, M., and G. Saxton. 2010. "The Dynamics of Ethnonationalist Contention." *British Journal of Political Science* 40 (2): 305–331. https://doi.org/10.1017/S0007123410000013

Berman, C. E. 2021. "Policing the Organizational Threat in Morocco: Protest and Public Violence in Liberal Autocracies." *American Journal of Political Science* 65 (3): 733–754. https://doi.org/10.1111/ajps.12565

Bizyukov, P., and J. M. Dollbaum. 2021. "Using Protest Event Analysis to Study Labour Conflict in Authoritarian Regimes: The Monitoring of Labour Protest Dataset." *Global Social Policy* 21 (1): 148–152. https://doi.org/10.1177/1468018121996076

Bodrunova, S. S., A. Litvinenko, and K. Nigmatullina. 2021. "Who Is the Censor? Self-Censorship of Russian Journalists in Professional Routines and Social Networking." *Journalism* 22 (12): 2919–2937. https://doi.org/10.1177/1464884920941965

Boeschoten, L., J. Ausloos, J. E. Möller, T. Araujo, and D. L. Oberski. 2022. "A Framework for Privacy Preserving Digital Trace Data Collection through Data Donation." *Computational Communication Research* 4 (2): 388–423. https://doi.org/10.5117/CCR2022.2.002.BOES

Borbáth, E., and S. Hutter. 2021. "Protesting Parties in Europe: A Comparative Analysis." *Party Politics* 27 (5): 896–908. https://doi.org/10.1177/1354068820908023

Boschee, E., J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward. 2015. "ICEWS Coded Event Data." *Harvard Dataverse*, V37. https://doi.org/10.7910/DVN/28075

Cha, M., H. Haddadi, F. Benevenuto, and K. Gummadi. 2010. "Measuring User Influence in Twitter: The Million Follower Fallacy." *Proceedings of the International AAAI Conference on Web and Social Media* 4 (1): 10–17. https://doi.org/10.1609/icwsm.v4i1.14033

Chang, C. H., N. R. Deshmukh, P. R. Armsworth, and Y. J. Masuda. 2023. "Environmental Users Abandoned Twitter After Musk Takeover." *Trends in Ecology and Evolution* 38 (10): 893–895. https://doi.org/10.1016/j.tree.2023.07.002

Chang, C., and M. Manion. 2021. "Political Self-Censorship in Authoritarian States: The Spatial-Temporal Dimension of Trouble." *Comparative Political Studies* 54 (8): 1362–1392. https://doi.org/10.1177/0010414021989762

Clarke, K. 2023. "Which Protests Count? Coverage Bias in Middle East Event Datasets." *Mediterranean Politics* 28 (2): 302–328. https://doi.org/10.1080/13629395.2021.1957577

Croicu, M., and N. B. Weidmann. 2015. "Improving the Selection of News Reports for Event Coding Using Ensemble Classification." *Research and Politics* 2 (4). https://doi.org/10.1177/2053168015615596
Davenport, C. 2009. *Media Bias, Perspective, and State Repression: The Black Panther Party*. New York: Cambridge University Press.
Dehghan, E., and S. Glazunova. 2021. "'Fake News' Discourses: An Exploration of Russian and Persian Tweets." *Journal of Language and Politics* 20 (5): 741–760. https://doi.org/10.1075/jlp.21032.deh
Denisova, A. 2017. "Democracy, Protest, and Public Sphere in Russia After the 2011–2012 Anti-Government Protests: Digital Media at Stake." *Media, Culture, and Society* 39 (7): 976–994.
Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Proceedings of the 2019 Conference Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019. (Association for Computational Linguistics), 4171–4186. https://doi.org/10.18653/v1/N19-1423
Deza, M. M., and E. Deza. 2009. *Encyclopedia of Distances*. Heidelberg: Springer.
Dollbaum, J. M. 2021. "Protest Event Analysis Under Conditions of Limited Press Freedom: Comparing Data Sources." *Media and Communication* 9 (4): 104–115. https://doi.org/10.17645/mac.v9i4.4217
Dovbysh, O., and O. Mukhametov. 2020. "State Information Contracts: The Economic Leverage of Regional Media Control in Russia." *Demokratizatsiya: The Journal of Post-Soviet Democratization* 28 (3): 367–391.
Dubina, V., and A. Arkhipova. 2023. "'No Wobble': Silent Protest in Contemporary Russia." *Russian Analytical Digest*, 291 : 8–11. https://doi.org/10.3929/ethz-b-000595208
Earl, J., A. Martin, J. D. McCarthy, and S. A. Soule. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology* 30:65–80. https://doi.org/10.1146/annurev.soc.30.012703.110603
Earl, J., S. A. Soule, and J. D. McCarthy. 2003. "Protest Under Fire? Explaining the Policing of Protest." *American Sociological Review* 68 (4): 581–606. https://doi.org/10.1177/000312240306800405
Fiesler, C., and B. Dym. 2020. "Moving Across Lands: Online Platform Migration in Fandom Communities." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW1): 1–25. https://doi.org/10.1145/3392847
Fisher, D. R., K. T. Andrews, N. Caren, E. Chenoweth, M. T. Heaney, T. Leung, L. N. Perkins, and J. Pressman. 2019. "The Science of Contemporary Street Protest: New Efforts in the United States." *Science Advances* 5 (10): eaaw5461. https://doi.org/10.1126/sciadv.aaw5461
Garrido-Merchán, Eduardo C. Gozalo-Brizuela Roberto, and Santiago González-Carvajal. 2023. "Comparing BERT Against Traditional Machine Learning Text Classification." *Journal of Computational and Cognitive Engineering* 2 (4): 352–356. https://doi.org/10.47852/bonviewJCCE3202838
Goncharenko, G., and I. Khadaroo. 2020. "Disciplining Human Rights Organisations Through an Accounting Regulation: A Case of the 'Foreign Agents' Law in Russia." *Critical Perspectives on Accounting* 72:102129. https://doi.org/10.1016/j.cpa.2019.102129
Hale, H. E. 2018. "How Crimea Pays: Media, Rallying 'Round the Flag, and Authoritarian Support." *Comparative Politics* 50 (3): 369–391. https://doi.org/10.5129/001041518822704953
Halterman, A., B. E. Bagozzi, A. Beger, P. Schrodt, and G. Scraborough. 2023. PLOVER and POLECAT: A New Political Event Ontology and Dataset. https://doi.org/10.31235/osf.io/rm5dw
Hanna, A. 2017. "MPEDS: Automating the Generation of Protest Event Data. SocArXiv preprint (Xuquv_v1). https://doi.org/10.31235/osf.io/xuqmv
He, L., B. Agard, and M. Trépanier. 2020. "A Classification of Public Transit Users with Smart Card Data Based on Time Series Distance Metrics and a Hierarchical Clustering Method." *Transportmetrica A: Transport Science* 16 (1): 56–75. https://doi.org/10.1080/23249935.2018.1479722
Hoffmann, M., F. G. Santos, C. Neumayer, and D. Mercea. 2022. "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. https://doi.org/10.1080/19312458.2022.2128099

Hopp, F. R., J. Schaffer, J. T. Fisher, and R. Weber. 2019. "ICORE: The GDELT Interface for the Advancement of Communication Research." *Computational Communication Research* 1 (1): 13–44. https://doi.org/10.5117/CCR2019.1.002.HOPP

Hutter, S. 2014. "Protest Event Analysis and Its Offspring." In *Methodological Practices in Social Movement Research*, edited by D. della Porta, 335–367. Oxford: Oxford University Press.

Hutter, S. 2019. "Exploring the Full Conceptual Potential of Protest Event Analysis." *Sociological Methodology* 49 (1): 58–63. https://doi.org/10.1177/0081175019860239

Isotani, H., H. Washizaki, Y. Fukazawa, T. Nomoto, S. Ouji, and S. Saito. 2021. "Duplicate Bug Report Detection by Using Sentence Embedding and Fine-Tuning." In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, New York: 535–544. Institute of Electrical and Electronic Engineers. https://doi.org/10.1109/ICSME52107.2021.00054

Johnson, E. W., J. P. Schreiner, and J. Agnone. 2016. "The Effect of New York Times Event Coding Techniques on Social Movement Analyses of Protest Data." In *Narratives of Identity in Social Movements, Conflicts, and Change*, edited by L.E. Hancock, 263–291. Bingley: Emerald Group.

King, B. G., and S. A. Soule. 2007. "Social Movements as Extra-Institutional Entrepreneurs: The Effect of Protests on Stock Price Returns." *Administrative Science Quarterly* 52 (3): 413–442. https://doi.org/10.2189/asqu.52.3.413

Koesel, K. J., and V. J. Bunce. 2012. "Putin, Popular Protests, and Political Trajectories in Russia: A Comparative Perspective." *Post-Soviet Affairs* 28 (4): 403–423. https://doi.org/10.2747/1060-586X.28.4.403

Koopmans, R., and P. Statham. 1999. "Political Claims Analysis: Integrating Protest Event and Political Discourse Approaches." *Mobilization: An International Quarterly* 4 (2): 203–221. https://doi.org/10.17813/maiq.4.2.d75933706071l6756

Kriesi, H., R. Koopmans, J. W. Duyvendak, and M. G. Giugni. 1995. *New Social Movements in Western Europe: A Comparative Analysis*. London: Routledge.

Kriesi, H., and I.-E. Oana. 2022. "Protest in Unlikely Times: Dynamics of Collective Mobilization in Europe During the COVID-19 Crisis." *Journal of European Public Policy* 30 (4): 740–765. https://doi.org/10.1080/13501763.2022.2140819

Kuratov, Y., and M. Arkhipov. 2019. "Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language". arXiv preprint (arXiv:1905.07213). https://doi.org/10.48550/arXiv.1905.07213

Lamberova, N., and K. Sonin. 2023. "Information Manipulation and Repression: A Theory and Evidence from the COVID Response in Russia." University of Chicago, Becker Friedman Institute for Economics Working Paper no. 2022-101.

Lankina, T., and R. Skovoroda. 2017. "Regional Protest and Electoral Fraud: Evidence from Analysis of New Data on Russian Protest." *East European Politics* 33 (2): 253–274. https://doi.org/10.1080/21599165.2016.1261018

Lankina, T., and K. Tertytchnaya. 2019. "Protest in Electoral Autocracies: A New Dataset." *Post-Soviet Affairs* 36 (1): 20–36. https://doi.org/10.1080/1060586X.2019.1656039

Leetaru, K., and P. A. Schrodt. 2013. "GDELT: Global Data on Events, Location, and Tone, 1979–2012." *ISA Annual Convention* 2 (4): 1–49.

Lewis, D. G. 2020. *Russia's New Authoritarianism: Putin and the Politics of Order*. Edinburgh: Edinburgh University Press.

Lipman, M. 2009. "Media Manipulation and Political Control in Russia." Carnegie Endowment for International Peace. Accessed November 27, 2025. https://carnegieendowment.org/posts/2009/02/media-manipulation-and-political-control-in-russia?lang=en.

Lipman, M. 2016. "At the Turning Point to Repression: Why There Are More and More 'Undesirable Elements' in Russia." *Russian Politics and Law* 54 (4): 341–350. https://doi.org/10.1080/10611940.2016.1207468

Littman, J., D. Chudnov, D. Kerchner, C. Peterson, Y. Tan, R. Trent, R. Vij, and L. Wrubel. 2018. "API-Based Social Media Collecting as a Form of Web Archiving." *International Journal on Digital Libraries* 19 (1): 21–38. https://doi.org/10.1007/s00799-016-0201-7

Lorenzini, J., H. Kriesi, P. Makarov, and B. Wüest. 2022. "Protest Event Analysis: Developing a Semiautomated NLP Approach." *American Behavioral Scientist* 66 (5): 555–577. https://doi.org/10.1177/00027642211021650

Makarov, P., J. Lorenzini, K. Rothenhäusler, and B. Wüest. 2015. "Towards Automated Protest Event Analysis." Zurich Open Repository and Archive (ZORA). Accessed November 27, 2025. https://www.zora.uzh.ch/entities/publication/790489d8-ccef-4cc6-a31c-e0279eb97115 .

Mamaev, B. 2024. *The Evolution of Authoritarianism and Contentious Action in Russia*. Cambridge: Cambridge University Press.

McAdam, D. 1999. *Political Process and the Development of Black Insurgency, 1930–1970*. Chicago: University of Chicago Press.

McInnes, L., J. Healy, J. Melville, and L. Großberger. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv preprint (arXiv:1802.03426). *Journal of Open Source Software* 3 (29): 861. https://doi.org/10.21105/joss.00861

Mejias, U. A., and N. E. Vokuev. 2017. "Disinformation and the Media: The Case of Russia and Ukraine." *Media, Culture, and Society* 39 (7): 1027–1042. https://doi.org/10.1177/0163443716686672

Mueller, C. 1997. "Media Measurement Models of Protest Event Data." *Mobilization: An International Quarterly* 2 (2): 165–184. https://doi.org/10.17813/maiq.2.2.n043476m01q7463u

Nam, T. 2006. "What You Use Matters: Coding Protest Data." *PS: Political Science and Politics* 39 (2): 281–287. https://doi.org/10.1017/S104909650606046X

Northcott, T., K. Sievert, C. Russell, A. Obeid, D. Angus, and C. Parker. 2025. "Unhealthy Food Advertising on Social Media: Policy Lessons from the Australian Ad Observatory." *Health Promotion International* 40 (2): daae192. https://doi.org/10.1093/heapro/daae192

Ohme, J., T. Araujo, L. Boeschoten, D. Freelon, N. Ram, B. B. Reeves, and T. N. Robinson. 2024. "Digital Trace Data Collection for Social Media Effects Research: APIS, Data Donation, and (Screen) Tracking." *Communication Methods and Measures* 18 (2): 124–141. https://doi.org/10.1080/19312458.2023.2181319

Oliver, P. E., J. Cadena-Roa, and K. D. Strawn. 2003. "Emerging Trends in the Study of Protest and Social Movements." *Research in Political Sociology* 12 (1): 213–244.

Olzak, S. 1989. "Analysis of Events in the Study of Collective Action." *Annual Review of Sociology* 15:119–141. https://doi.org/10.1146/annurev.so.15.080189.001003

Olzak, S. 1992. *The Dynamics of Ethnic Competition and Conflict*. Stanford, CA: Stanford University Press.

Ong, E. 2021. "Online Repression and Self-Censorship: Evidence from Southeast Asia." *Government and Opposition* 56 (1): 141–162. https://doi.org/10.1017/gov.2019.18

OVD-Info. 2024. "Persecution of the Anti-War Movement Report. Two Years of Russia's Full-Scale Invasion of Ukraine." Accessed February 28, 2024. https://ovd.info/en/persecution-anti-war-movement-report-two-years-russias-full-scale-invasion-ukraine.

Pan, J. 2017. "How Market Dynamics of Domestic and Foreign Social Media Firms Shape Strategies of Internet Censorship." *Problems of Post-Communism* 64 (3–4): 167–188. https://doi.org/10.1080/10758216.2016.1181525

Pealat, C., G. Bouleux, and V. Cheutet. 2021. "Improved Time-Series Clustering with UMAP Dimension Reduction Method." In *2020 25th International Conference on Pattern Recognition (ICPR)* (IEEE), 5658–5665. https://doi.org/10.1109/ICPR48806.2021.9412261

Poupin, P. 2021. "Social Media and State Repression: The Case of VKontakte and the Anti-Garbage Protest in Shies, in Far Northern Russia." *First Monday* 26 (5). https://doi.org/10.5210/fm.v26i5.11711

Raleigh, C., A. Linke, H. Hegre, and J. Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset." *Journal of Peace Research* 47 (5): 651–660. https://doi.org/10.1177/0022343310378914

Ramesh, R., R. S. Raman, M. Bernhard, V. Ongkowijaya, L. Evdokimov, A. Edmundson, S. Sprecher, M. Ikram, and R. Ensafi. 2020. "Decentralized Control: A Case Study of Russia." In *Network and Distributed Systems Security (NDSS) Symposium 2020* (NDSS). Accessed November 27, 2025. https://doi.org/10.14722/ndss.2020.23098

Reimers, N., and I. Gurevych. 2019:. "Sentence-BERT: Sentence Embeddings Using Siamese bert-Networks." arXiv preprint (arXiv:1908.10084). https://doi.org/10.48550/arXiv.1908.10084

Riquelme, F., and P. González-Cantergiani. 2016. "Measuring User Influence on Twitter: A Survey." *Information Processing and Management* 52 (5): 949–975. https://doi.org/10.1016/j.ipm.2016.04.003

Roudakova, N. 2017. *Losing Pravda: Ethics and the Press in Post-Truth Russia*. Cambridge: Cambridge University Press.

Salehyan, I., C. S. Hendrix, J. Hamner, C. Case, C. Linebarger, E. Stull, and J. Williams. 2012. "Social Conflict in Africa: A New Database." *International Interactions* 38 (4): 503–511. https://doi.org/10.1080/03050629.2012.697426

Sapkota, R., S. Raza, and M. Karkee. 2025. "Comprehensive Analysis of Transparency and Accessibility of ChatGPT, DeepSeek, and Other SOTA Large Language Models." arXiv preprint (arXiv:2502.18505). https://doi.org/10.48550/arXiv.2502.18505

Scarborough, G. I., B. E. Bagozzi, A. Beger, J. Berrie, A. Halterman, P. A. Schrodt, and J. Spivey. 2023. "POLECAT Weekly Data." Harvard Dataverse. Accessed November 27, 2025. https://dataverse.harvard.edu/dataset.xhtml?persistentId.

Schimpfössl, E., and I. Yablokov. 2020. "Post-Socialist Self-Censorship: Russia, Hungary, and Latvia." *European Journal of Communication* 35 (1): 29–45. https://doi.org/10.1177/0267323119897797

Sousa, S., and R. Kern. 2023. "How to Keep Text Private? A Systematic Review of Deep Learning Methods for Privacy-Preserving Natural Language Processing." *Artificial Intelligence Review* 56 (2): 1427–1492. https://doi.org/10.1007/s10462-022-10204-6

Strawn, K. 2008. "Validity and Media-Derived Protest Event Data: Examining Relative Coverage Tendencies in Mexican News Media." *Mobilization: An International Quarterly* 13 (2): 147–164. https://doi.org/10.17813/maiq.13.2.b3j3p1104244u073

Sundberg, R., and E. Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4): 523–532. https://doi.org/10.1177/0022343313484347

Tarrow, S. G. 1989. *Democracy and Disorder: Protest and Politics in Italy, 1965–1975*. Oxford: Clarendon Press.

Trezza, D. 2023. "To Scrape or Not to Scrape, This Is Dilemma. The Post-API Scenario and Implications on Digital Research." *Frontiers in Sociology* 8:1145038. https://doi.org/10.3389/fsoc.2023.1145038

Venturini, T., and R. Rogers. 2019. "'API-Based Research' Or How Can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach." *Digital Journalism* 7 (4): 532–540. https://doi.org/10.1080/21670811.2019.1591927

Ward, M. D., A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford. 2013. "Comparing GDELT and ICEWS Event Data." *Analysis* 21 (1): 267–297.

Weidmann, N. B., and E. G. Rød. 2019a. *The Internet and Political Protest in Autocracies*. New York: Oxford University Press.

Weidmann, N. B., and E. G. Rød. 2019b. "Coding Protest Events in Autocracies." In *The Internet and Political Protest in Autocracies*, edited by Nils B. Weidmann and Espen Geelmuyden Rød, 39–60. New York: Oxford University Press.

Wiedemann, G., J. M. Dollbaum, S. Haunss, P. Daphi, and L. D. Meier. 2022. "A Generalized Approach to Protest Event Detection in German Local News." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC. (European Language Resources Association)* edited by Calzolari Nicoletta, Béchet Frédéric, Blache Philippe, Choukri Khalid, Cieri Christopher, Declerck Thierry, Goggi Sara, Isahara Hitoshi, Maegaard Bente, Mariani Joseph, Mazo Hélène, Odijk Jan, Piperidis Stelios, 3883–3891. https://aclanthology.org/2022.lrec-1.413/

Wüest, B., K. Rothenhäusler, and S. Hutter. 2013. "Using Computational Linguistics to Enhance Protest Event Analysis." Accessed November 27, 2025. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2286769 .

Yin, Z., D. Li, and D. W. Goldberg. 2023. "Is Chat-GPT a Game Changer for Geocoding—A Benchmark for Geocoding Address Parsing Techniques." arXiv preprint (arXiv:2310.14360).

Zhang, H., and J. Pan. 2019. "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology* 49 (1): 1–57. https://doi.org/10.1177/0081175019860244

Zhang, Z., Y. Zhao, H. Gao, and M. Hu. 2024. "Linkner: Linking Local Named Entity Recognition Models to Large Language Models Using Uncertainty." arXiv preprint (arXiv:2402.10573). https://doi.org/10.48550/arXiv.2402.10573