Plover and Polecat: A New Political Event Ontology and Dataset*

Andrew Halterman[†], Benjamin E. Bagozzi[‡], Andreas Beger[§], Philip A. Schrodt[¶], Grace I. Scarborough[‖]

April 1, 2023

Abstract

Polecat is a new global political event dataset intended to serve as the successor to the dataset produced by the DARPA Integrated Conflict Early Warning System (ICEWS) project. Polecat's event data are machine coded from millions of multi-language international news reports and will soon cover the period 2010-to-present. These data are generated using the Next Generation Event Coder (NGEC), a new automated coder that replaces the use of extensive (and difficult to update) dictionaries with a more flexible set of expert annotations of an event's characteristics. In contrast to existing automated event coders, it uses a combination of NLP tools, transformer-based neural networks, and actor information sourced from Wikipedia. Polecat's event data are based on an event-mode-context ontology, the Political Language Ontology for Verifiable Event Records (Plover), that replaces the older CAMEO ontology used in past datasets such as ICEWS and Phoenix. These innovations offer substantial improvements in the scope and accuracy of political event data in terms of the what, how, why, where, and when of domestic and international interactions. After detailing Plover and Polecat, we illustrate the innovations and improvements through a preliminary comparison to the existing-ICEWS event data system.

*Funding for Plover was initially provided in part by the U.S. National Science Foundation award SBE-1539302, "RIDIR: Modernizing Political Event Data for Big Data Social Science Research". Contributors to the development of Plover include, in alphabetical order, Benjamin Bagozzi, Andreas Beger, John Beieler, Liz Boschee, Patrick Brandt, Andrew Halterman, Jill Irvine, Jennifer Holmes, Javier Osorio, Tom Parris, Grace Scarborough, and Philip Schrodt. All subsequent refinements to Plover, development of NGEC, the Polecat database, and this paper were sponsored by the Political Instability Task Force (PITF). The PITF is funded by the Central Intelligence Agency. The views expressed in this paper are the authors' alone and do not represent the views of the U.S. Government. We both acknowledge and are deeply grateful for the work on the Leidos side of this development by John Berrie, Heather Carter, Kathleen Krupenevich, and Jevon Spivey.

[†]Michigan State University, ahalterman0@gmail.com

[‡]University of Delaware, bagozzib@udel.edu

[§]Basil Analytics, adbeger@gmail.com

[¶]Parus Analytics, schrodt735@gmail.com

[‖]Leidos, grace.i.scarborough@leidos.com

Introduction

Text-derived event data, or structured records of who-did-what-to-whom, are an important source of data for scholars of international relations and comparative politics (e.g., Goldstein and Freeman, 1990; Reuveny and Kang, 1996; Colaresi, 2004; Chiba and Gleditsch, 2017; Weschle, 2018; Blair and Sambanis, 2020; Kibris, 2021). However, existing machine-coded event datasets such as the Integrated Crisis Early Warning System (ICEWS; O'Brien, 2010; Boschee et al., 2015) and Phoenix Historical Dataset (Althaus et al., 2019) have several major limitations: they use a Conflict and Mediation Event Observations (CAMEO) coding ontology (Schrodt, Gerner and Yilmaz, 2009) that is overly complex and rigid in how it represents events, they rely on coding software that is opaque and difficult to update, they use dictionaries that have limited coverage and that go out of date, and they often lack sufficient validation against other event datasets. Each of these limits the usefulness of machine coded event data. At the same time, the global news sources available for coding near-real-time event data have completely outpaced any projects attempting to code these news sources with only human coders. Hence, and despite its imperfections, event data sets that are regularly and systematically updated, such as the ICEWS data on the open Dataverse site, will out of necessity need to use machine coding moving forward.

This paper introduces the POLitical Event Classification, Attributes, and Types (POLECAT) dataset—a new global event dataset that addresses these needs. POLECAT uses a novel coding ontology known as the Political Language Ontology for Verifiable Event Records (PLOVER) to replace CAMEO. The new coder that we employ to generate POLECAT implements PLOVER in conjunction with a variety of recently developed tools for automated event coding such as transformer-based large language model neural networks and "dictionary free" Wikipedia lookups. Together we contend that these innovations produce more accurate and useful data than previous large-scale event data projects such as ICEWS. The paper begins by describing PLOVER, our newly proposed coding ontology, before detailing POLECAT and its generation process. Following this, we compare POLECAT to ICEWS.

We find that our new Plover ontology and Polecat dataset have several advantages over existing systems. In contrast to CAMEO, the ontology it replaces, Plover has a relatively smaller number of event types (16, compared to over 250), and now distinguishes between the "how" of an event and the "why" of the event's broader context with additional fields. This novel ontology underpins Polecat, which is coded coded from English and machine translated stories and will soon cover the period 2010 to present. In contrast to extant automated event coding systems such as TABARI (Schrodt, 2011) and PETRARCH2 (Norris, Schrodt and Beieler, 2017), the coder used to generate Polecat does not rely on dictionaries, which are labor intensive to update and frequently miss events and actors. The coder is also modular and employs machine learning models that are easy to update as classifiers improve and more training data become available. The initial evaluations of Polecat reported in this paper show that it is comparable to ICEWS in event coverage of relevant domestic and international events although it codes more events overall, provides more conservative (and hence potentially more accurate) geolocations than does ICEWS, and offers additional contextual information on each recorded event than is available in ICEWS.

Background and Motivation

The concept of political event data originated in the academic quantitative international relations community in the mid-1960s. While a number of projects produced some event data, often for specialized applications, eventually two coding frameworks dominated the production of general-purpose event data sets: Charles McClelland's WEIS (McClelland, 1967, 1976) and the Conflict and Peace Data Bank (COPDAB) developed by Edward Azar (Azar and Sloan, 1975; Azar, 1980, 1982). Both were created during the Cold War and assumed a "Westphalian-Clausewitzian" political world in which sovereign states reacted to each other primarily through official diplomacy and military threats. Consequently these coding systems proved less than optimal for dealing with post-Cold-War issues such as ethnic conflict, low-intensity violence, internal conflict and repression, and multilateral intervention.

These early political event datasets relied upon extensive training (in both codebooks and person-hours) of students for coding events from news reports. In the 1990s, this gave way to two key advances in event data coding. First, the source material for event data projects largely shifted away from domestic U.S. newspapers and towards international news wire services. With their broader audiences and lower space constraints, these news wires generally provided more extensive and systematic international coverage of political interactions than did domestic oriented print newspapers, especially for more peripheral countries and actions. Second, and partly thanks to the machine-readability of news wires, the coding of event data became increasingly automated over the course of the 1990s. The systems developed for machine coding event data coding during this period (Gerner et al., 1994; Bond et al., 1997) used shallow parsing techniques to identify entities and actions within the lede sentences of news wire reports. Automated coding systems then cross-referenced these against extensive vocabulary dictionaries. Dictionary matches were then populated into event-level datasets along with corresponding event dates that were separately recovered from the original news wire report's meta data.

These innovations led to the development of a number of new event data ontologies in the early 2000s (Bond et al., 2003; King and Lowe, 2003; Schrodt, Gerner and Yilmaz, 2009). Most notably, the CAMEO framework (Schrodt, Gerner and Yilmaz, 2009) extended WEIS to support an NSF-funded project at the University of Kansas. This was primarily undertaken to study interstate conflict mediation, not as a means of creating a new general-purpose event ontology. Nonetheless, it was gradually adopted as a 'default coding scheme,' notably for the DARPA-funded Integrated Conflict Early Warning System (ICEWS) project (O'Brien, 2010), because it corrected some of the long-recognized ambiguities in WEIS and COPDAB, and was explicitly designed both for automated coding and for the detailed coding of sub-state actors. It was continued in the widely-used public ICEWS data (Boschee et al., 2015) coded using the BBN SERIF/ACCENT coder, with BBN doing considerable additional work on various details of the system. Alongside these innovations, a variety of open-source

14			PROTEST
	141		Demonstrate or rally
		1411	Demonstrate or rally for leadership change
		1412	Demonstrate or rally for policy change
	142		Conduct hunger strike
		1421	Conduct hunger strike for leadership change
		1422	Conduct hunger strike for policy change

Table 1: Example of CAMEO's hierarchical coding scheme. Note that many of the lowest level categories have few or no associated verb patterns, meaning that they would never be coded in practice.

event datasets using the CAMEO framework and the PETRARCH(2) coding system were released, including the Phoenix Near-Real-Time Dataset (Open Event Data Alliance, 2015), TERRIER (Grant et al., 2019), and the Phoenix Historical Dataset (Althaus et al., 2019).

As machine-coded event data came into wider use in the 2010s, however, several problems with CAMEO became apparent (Beieler et al., 2016). First, almost all applications of CAMEO event data aggregated to either the 20 coarser "cue categories" or the even more general "quad categories."[1] Virtually no one used all 260 event categories. Nonetheless, users unfamiliar with the data generating process for automated event coding sometimes assumed every (two-, three-, and four-digit) CAMEO event category—such as those presented for CAMEO's PROTEST category in Table 1—had been equally well implemented. Second, the complexity of CAMEO made it almost impossible to generate a comprehensive set of "gold standard records" and human coders had difficulty agreeing on how to consistently distinguish many of the subcategories.[2] Third, newer coding systems provided information such as geolocation and named-entity extraction beyond the original date-source-target-event format and there was no standard for how to include these in event coding pipelines employing CAMEO. Fourth, the continuing emphasis on coding substate activities demonstrated the need for either revised event categories or new event contexts such as natural disasters,

[1] Verbal cooperation, material cooperation, verbal conflict, and material conflict, with these usually defined exclusively using the 2-digit categories.

[2] This became particularly apparent as efforts were made to implement CAMEO in Spanish and Arabic (Halterman et al., 2018; Osorio et al., 2019).

elections, and parliamentary behaviors. Yet, at the same time, the hurdles associated with creating new dictionaries under the CAMEO framework were prohibitively costly (Radford, 2021).

PLOVER

To address the challenges outlined above, an informal group of academic, government and private sector producers and users of event data met and circulated drafts during fall 2016 to develop a new, simplified, and more flexible event data ontology to replace CAMEO, which ultimately became PLOVER. Additional extensive work was done in 2021 as PLOVER was adopted by the Political Instability Task Force (PITF) to replace CAMEO in its new coder for the dataset introduced below as POLECAT.

Because the PLOVER event categories are generally a simplification of CAMEO, it is relatively easy to splice existing CAMEO data sets to PLOVER equivalents by simply collapsing CAMEO's two- and three-digit categories. Similar aggregate linkages between CAMEO and PLOVER can likewise be easily implemented using the quad categories referenced above. A scaled set of PLOVER scores (described further below) are also designed for splicing with time series data generated from the CAMEO Goldstein scores (Goldstein, 1992).

Compared to the existing CAMEO manuals (Schrodt, 2012; Boschee et al., 2015)—though curiously, consistent with the public documentation for WEIS and COPAB (Azar, 2009; McClelland, 2006)—we provide only general guidance on the content of the various `event types`, `modes`, and `contexts` in PLOVER. With the current state of automated natural language processing, any future automated coding system will almost certainly be implemented using machine learning systems trained on a labeled set of news texts and those training cases effectively are the detailed examples. This differs from the older systems that classified events using dictionaries that were abstracted, by human developers, from the texts. We accompany the final POLECAT data with annotated synthetic training cases that are free of intellectual property constraints (Halterman, 2022).[3]

[3]There is also a small set of cases extracted from the CAMEO manual that is available on the

A major innovation of PLOVER is the reallocation of the information contained within the hierarchical 3- and 4-digit categories of CAMEO into three components: `event-mode-context`. These components can be generally interpreted as capturing the `what-how-why` characteristics of any given event. We anticipate at least five advantages to this approach. First, PLOVER's `what-how-why` components are now distinct, whereas various CAMEO subcategories inconsistently used the *how* and *why* to distinguish between subcategories. Second, because `context` can be applied to any event category and, where relevant, any `mode`, PLOVER has far more combinations of codes for describing events than the fixed hierarchy of CAMEO. Third, the "`what-how-why`" formulation increases the ability of general machine-learning classifiers—as distinct from the older customized dictionary-based parser/coders—to assign `mode` and `context` compared to their ability to assign CAMEO subcategories. Fourth, separate "`what-how-why`" components are *much* easier for humans to code than the hierarchical structure of CAMEO because a human coder can independently hold most of the relevant categories in working memory when coding.[4] Fifth, because the words used to differentiate `mode` and `context` are generally very basic, translations of the coding protocols into languages other than English is likely to be easier than translating CAMEO's subcategory description.

While `context` and (especially) `mode` will often have only a single value for a given event record, in some instances multiple values will be appropriate and allowed. Both `context` and `mode` are optional fields. If no existing values seem appropriate, the relevant `context` and `mode` field is left null. This can be contrasted with the use of "NA" for `mode` if an `event type` itself has no allowable `modes` based on PLOVER. With these caveats in mind, we now turn to a more detailed discussion of PLOVER's `event types`, `contexts`, and `modes`. Full summaries and tables—with associated definitions and coding rules—of each individual `event type`, `context`, and `mode` appear in the Supplemental Appendix.

PLOVER GitHub site (`https://github.com/openeventdata/PLOVER`) but it is not sufficient for training a system, at least with contemporary technologies.

[4]More generally, `event-mode-context` coding uses words, not numerical codes, so coders will probably be using the parts of the brain (Broca's area) which are specialized for processing words. No known specialized cognitive facility exists for handling some 250 2-to-4-digit codes.

`Event types` correspond to the `what` of a given event between a particular actor and recipient (or a set of actors and recipients). `Event types` include, for example, the initiation of a military retreat from a particular territory by an occupying force or the verbal threat of violence against a government by an armed non-state organization. PLOVER utilizes a set of 16 overarching `event types` for the classification of events into distinct (verbal or material) cooperative or hostile `event types`. These `event types` were primarily developed to improve upon the 20 "high-level" event cue-categories used in CAMEO and thus in earlier event datasets such as ICEWS.[5] Finer grained event categories common to CAMEO and event datasets such as ICEWS, TERRIER, and the Phoenix Historical Dataset are then handled via `modes` as opposed to separate event categories.

Under the PLOVER event data ontology, `modes` correspond to the `how` of a given event. For example, within the PROTEST `event type` in PLOVER, there are separate available `modes` that indicate whether an identified protest event was undertaken via a demonstration, a riot, a strike, a hunger strike, a boycott, or an effort to obstruct access to a particular location. `Modes` are always specific to PLOVER's individual `event types` and five of PLOVER's 16 `event types` do not have any associated `modes`. As noted above, the closest correspondence between PLOVER's `modes` and the CAMEO event data ontology lies in CAMEO's 3- and 4-digit categories. As can be seen in Table 1, some low-level CAMEO categories contained "how" information and others contained "why" information. Because the latter categories were at times only sparsely or inaccurately coded in practice, PLOVER developed a more accurate and realistic set of `modes` for each relevant `event type`, following a series of expert discussions and human coding trials.

PLOVER's `contexts` record the issue area(s) surrounding a particular event. As such, the `context` field re-introduces, albeit in a greatly extended form, a concept found in the original COPDAB data (but absent from WEIS and hence CAMEO) which allows for a distinction

[5] To go from 20 to 16 primary categories, CAMEO's two purely verbal MAKE PUBLIC STATEMENT and APPEAL were dropped (most analyses ignored them anyway), and ASSAULT, FIGHT, and ENGAGE IN UNCONVENTIONAL MASS VIOLENCE were combined into the single ASSAULT with their distinctions now delineated by `modes`.

in the event record between a meeting dealing with military issues and a meeting dealing with economic issues, for example. Human analysts naturally incorporate this information in their reading of an article. Based on some initial experiments, we believe that with contemporary text classification algorithms this is relatively easy to implement. Whereas `modes` are specific to individual `event types`, `context` is coded at the story-level and can arise for any `event types` (and `modes`). As outlined in the Supplemental Appendix, there are currently 37 `contexts` in total, ranging from broader themes such as "military" and "economic" to more specific issue areas such as "migration," "illegal drugs," and "LGBT."

Because `contexts` are applied using document-level classifiers,[6] researchers/analysts should be careful when interpreting the meaning(s) of `contexts`. For example, an ASSAULT event with an "elections" `context` does not automatically imply electoral violence. It could be a news article about, for instance, violence in Afghanistan in the `context` of an article on Pakistani election politics. Thus, `context` is ideal for filtering events that arise against the backdrop of a certain issue or thematic area, but not for inferring that an event included a particular `context` within that event's `mode` of occurrence. As noted above, certain types of events, particularly general protests and meetings, will also have multiple `contexts`.

Beyond an event's `what-how-why`-components, the PLOVER ontology also made refinements to an event's `who`-component. The latter corresponds to the entities designated as initiating and—where applicable—receiving a particular event. First, and in order to reduce overlap with how the policy community uses the terms "source" and "target", PLOVER utilizes the term `actor` to refer to the entity or entities who initiated the event, and `recipient` to refer to the entity or entities to whom the event is directed, if this is clear. Importantly, `recipient` is optional for many events.

For PLOVER's `actor`- and `recipient`-specific "country name" fields, there are three

[6] Though note that, technically speaking, POLECAT considers a "document" to correspond to a news article's first 512 "word pieces" (Devlin et al., 2019) after preprocessing articles to remove datelines, embedded newslines, editorial processing notes, and other material not relevant to the content of the story: detailed information on this filtering can be found in Halterman et al. (2023). In the corpora we are working with, in over 70% of cases the entire story fits into this limit, and it is also consistent with several earlier event data sets, including most iterations of ICEWS, which only coded the first four to six sentences of a story.

groups of entities that can arise. The first is independent nation states,[7] which are identified via either full country name (the default) or a country's ISO 3166 3-letter code. Second, by exception, a small number of non-independent territories and dependencies with ISO 3166 codes—such as Hong Kong and Palestine—are also included. Third, a small number of international non-state entities such as non-governmental organizations (NGOs) and multi-national corporations (MNCs) can also be included. Additional details on each set of entities can be found in the Supplemental Appendix. Note that the POLECAT data below is implemented in a manner that retains (i.e., lists) multiple `actor` and/or `recipient` countries within a single event entry in the data, where applicable,[8] rather than misleadingly breaking these multi-`actor` and/or multi-`recipient` events into separate dyadic events.

Alongside an event's `actor` and `recipient` "country names" fields, PLOVER also reports each `actor` and `recipient`'s associated "sectors," where appropriate. CAMEO employed a hierarchical agent coding structure based on 3-character coding elements which allowed nearly unlimited complexity and, depending on the exact coding system, could be resolved down to the identity of individual groups or individuals. As with CAMEO's event codes, typically only the first two or three of these elements were ever used by researchers. ICEWS modified this framework somewhat, while preserving most of the sub-state differentiations as "sectors"—the terminology we've adopted here over the CAMEO/IDEA "agents" terminology—while also providing a very substantial amount of complexity at the sub-sector level. Relative to ICEWS, PLOVER's sectors revert back to a simplified version of the agents contained in CAMEO—encompassing 25 distinct sector categories in total. These sectors are fully listed in the Supplemental Appendix and encompass well known designations from past event datasets such as COP (police), GOV (government, and REB (rebel), as well as historically lesser-used designations such as REF (refugee), CRM (criminal), and CIV (civilian). For each identified `actor` and `recipient`, multiple sector codes are possible in PLOVER, though the tiered nature of PLOVER's sectors allow for the recovery of a "primary"

[7] As based upon the Correlates of War (COW) state system membership list (COW, 2017).

[8] E.g., as in the case of a multilateral meeting.

sector code separately from an `actor` or `recipient`'s full list of identified sectors.

Finally, the level of hostility or cooperation is an event attribute that researchers and analysts often wish to capture. CAMEO incorporated the Goldstein (1992) extension of the WEIS scale. Events were coded on a -10 to $+10$ scale, though the full range was not used. The PLOVER ontology provides comparably scaled conflict/cooperation scores for events. As described in the Supplemental Appendix, the PLOVER scores themselves were created by taking a rescaled average of two distinct approaches to translating Goldstein scores into each PLOVER `event type-mode` combination, alongside a small number of qualitative adjustments to these averages based upon subject matter expert input. PLOVER likewise provides equivalent event type categorizations for what came to be commonly known as quad categories in CAMEO (e.g., Chiba and Gleditsch, 2017; Beger, Morgan and Ward, 2021). These categories effectively allow researchers to aggregate all PLOVER events into more general themes of verbal cooperation, material cooperation, verbal conflict, and material conflict. Additional details on quad categories, PLOVER scores, and the similarities and differences between PLOVER and CAMEO appear in the Supplemental Appendix.

POLECAT

POLECAT is a new global political event dataset that leverages PLOVER and a series of new innovations for the machine coding of events and event attributes. POLECAT is designed to replace the longstanding ICEWS event dataset and hence encompasses domestic and international events across the entire globe. It does not include purely domestic U.S.-events. Above and beyond the PLOVER innovations outlined above, the POLECAT dataset offers four additional improvements over existing automated event datasets and coders.

First, the corpus of news texts used by POLECAT encompasses an especially diverse and extensive collection of politically relevant news sources. POLECAT's news corpus includes thousands of news(wire) sources for the 2010-present period with near real-time updating and weekly postings of new releases to Dataverse moving forward, as has been done with the

CAMEO-coded ICEWS data. Notably, and following the application of a series of specialized search strings for initial story filtering, POLECAT's corpus includes politically relevant articles written in English, Spanish, French, Portuguese, Arabic, Russian, and Chinese. This inclusion of non-English news articles helps to address a number of previously highlighted biases in international relations-focused text analysis (Windsor, 2022). For POLECAT these non-English articles are machine translated to English before coding. Together, the corresponding breadth and timeliness of POLECAT's news article inputs ensure a more extensive and timely set of coded political events than most other global event datasets currently available.

In comparison to ICEWS, specifically, we note that most of the existing ICEWS source texts—news articles from a wide variety of sources, though the majority are from the major international news agencies Reuters, Agence France Presse and BBC—date from the DARPA and immediate post-DARPA periods when the project was exclusively focused on Asia. This quite conspicuously meant that India was massively over-sampled, thanks to its ready availability of English-language news sources. This over-sampling has been corrected since PITF took over the ICEWS project. The ICEWS data and news sources were also inconsistent in terms of the presence of non-English sources and the quality of their translations: ICEWS gradually introduced Spanish sources, later extending to Portuguese, and French. Prior to about 2000, low-quality translation software was used, and it is likely that the patterns used in the dictionaries—for the most part developed on stories written, or at least edited, by native speakers of English—missed a great deal of these events. As alluded to above, POLECAT has added Chinese, Russian, and Arabic using Google's translation systems and is unlikely to run into this issue.

Second, Althaus, Peyton and Shalmon (2022) draw attention to the problem of historical events in machine coded event data. The authors point out that past automated event datasets such as ICEWS have no known mechanism for handling historical events. This often leads to coding historical discussions of events as contemporary events based upon a news article's dateline. For instance, following Boko Haram's 2014 kidnapping of girls from Chibok, Althaus, Peyton and Shalmon (2022) find that the number of coded events in ICEWS on subsequent anniversaries of the event is almost as great or sometimes greater than the number of coded events from the original kidnapping. POLECAT's coding pipeline includes a mechanism for extracting the date of the event from within the story itself, rather than relying on a news article's dateline as has been the case for many past automated event datasets. This allows POLECAT to more accurately distinguish between the date of publication and the reported date of the event itself. The POLECAT coding pipeline then uses a similar approach for automated event geolocation, rather than relying upon bylines for geolocation. This again helps POLECAT to avoid past automated event data coding issues with respect to inaccurately geolocating events to (e.g.) country capitals based solely on the location from which a news story was filed, or locations based on the organizational location of the news source, such as New York, London, or Teheran, which is a not-infrequent issue in the ICEWS data.

Third, rather than relying on a rule-based coder, POLECAT uses a flexible supervised machine learning framework for identifying and coding an event's `event type`, `mode`,`context`, `actor` and `recipient`. For each PLOVER category associated with each of the above attributes, expert coders labeled hundreds of positive and negative labels using an active learning-directed semi-random sample of politically relevant news articles. These human labels were utilized as inputs to POLECAT's supervised machine learning classifiers. Support vector machines (SVMs; Cortes and Vapnik, 1995) were used for coding `modes` and `contexts`. POLECAT next employed a transformer-based neural network model, specifically distilBERT (Sanh et al., 2019),[9] for the supervised classification of `event types`. The closely related RoBERTa (Liu et al., 2019) was then similarly leveraged within a "question answering" (QA) framework (FitzGerald et al., 2018; Du and Cardie, 2020)[10]—alongside a

[9]distilBERT and roBERTa are two reduced implementations of Google's BERT large language model, which was trained on 3.4-billion words of English text. All are open access and available through the HuggingFace system.

[10]In essence, these models take in a passage of text (e.g., "a car bombing in downtown Aleppo has killed

series of category-specific `event type` and `mode` tailored questions—for recovering each identified event's associated entities (e.g., `actor` and `recipient`). Together these steps ensure that POLECAT obtains levels of accuracy in event coding and entity resolution that notably outpace those obtained under prior machine learning systems, which themselves have long exceeded the accuracy of teams of human coders in large-scale projects sustained over time. These innovations furthermore *modularize* POLECAT's underlying coding framework so as to allow future coder updates to specific components. This is distinct from the more interconnected and hence rigid TABARI, PETRARCH(2), and BBN SERIF/ACCENT frameworks of past event data systems.

Fourth, rather than relying on difficult to maintain dictionaries for coding each `actor` and/or `recipient`'s relevant country and sector code(s), POLECAT utilizes automated entity lookups. Specifically, given an `actor` or `recipient` phrase obtained for a particular event from the QA process mentioned above, POLECAT automatically resolves the entity to its unique Wikipedia identifier. Wikipedia is frequently updated for current events, thus addressing the challenges of manually maintaining and updating CAMEO's actor and agents dictionaries. Wikipedia furthermore contains both the types of information found in CAMEO's customized role dictionaries[11] and provides an immediate reference to additional biographical and historical information. It also has a standard format for political biographies, allowing this lookup step to uniformly recover relevant `actor` or `recipient` country, sector, and date-range inputs for PLOVER's corresponding country and sector fields. This POLECAT innovation thereby does away with the need for actor dictionaries almost entirely, with the exception of a relatively small file that maps generic actor descriptions and office titles to their PLOVER codes. Together, these innovations significantly expand the capacity of POLECAT to capture relevant current and historical actors in near real-time.

several people") and asks a question in natural language (e.g., "Who was killed in the assault?"). The model then returns a phrase from the document that answers that specific question (e.g., "several people").

[11] Which following the end of the ICEWS-BBN work appear to have only been updated every six months or so, and this process seems to have been entirely manually done.

Coding Process

POLECAT's coding pipeline is based upon a new automated coder known as the Next Generation Event Coder (NGEC; Halterman et al., 2023). Some components of NGEC—and hence the POLECAT pipeline—require human labeled data as inputs to POLECAT's supervised machine learning models, whereas other NGEC-components are implemented in a fully automated fashion.[12] In the current subsection, we first discuss POLECAT's human labeling components and assessments. We then summarize the NGEC coding steps that ultimately produce POLECAT in further detail. Note that an even more expanded discussions of these various NGEC steps and inputs can be found in Halterman et al. (2023) and in the Supplemental Appendix.

The human labeling component to POLECAT involved approximately 10 expert human coders. Politically-relevant news stories were drawn from the ICEWS project's news corpus for a large cross-section of time spanning the past two decades. Stories were provided to coders in a semi-random fashion using active learning to continuously update selected stories for coding based upon an underlying machine learning model. After a series of training sessions, human coders were asked to label event `context` at the news story level. `Event types` were then labeled within individual news stories by underlining relevant passages of text. `Modes` were labeled at the story level conditional on a news story having been identified as containing a relevant `mode`'s overarching `event type`.[13] Finally, an event's entities (i.e., `actor`, `recipient`, date, and location) were individually highlighted within a set of news stories that had been previously identified as containing events. Each of these steps was repeated for every `context`, `event type`, and `mode` category. Each `context` and `event type` category ultimately received approximately 250 positive labels and 1000 negative labels, and each `mode` received between 50-200 positive labels and 500-1000 negative labels. Remote calibration meetings were held on a bi-weekly basis during human labeling to enhance coder

[12]Such as the Wikipedia lookup step mentioned immediately above.

[13]To ensure sufficient story samples for each `event type-mode` combination, this step (and the subsequent entity labeling step) utilized a combination of real news stories and synthetically generated news stories (Halterman, 2022).

accuracy and agreement.

Evaluations of coding accuracy were undertaken iteratively during classifier selection and tuning. These evaluations were at times conducted with all human labels treated as ground truth within the context of in-sample and out-of-sample supervised classifier performance. Separately, some evaluations also evaluated machine coder performance in relation to a gold-standard sample of nearly 1000 manually checked, double- or triple-human coded POLECAT events. For `mode` and `context`, SVMs were found to perform with notable precision and recall—generally in the 0.70-0.90 range when evaluated on a split sample of annotated stories. This is comparable to past event data projects and alternative supervised classifiers.[14] This favored SVMs for these event coding components. However, coding `event types` and event entities required more recently developed language models to ensure adequate coder accuracy. Using the evaluation samples mentioned above, four such approaches[15] were considered in terms of both accuracy and compute time. A fine-tuned RoBERTa question-answering model was found to perform best in coder accuracy and third-best in terms of compute time, leading it to be favored for POLECAT's attribute identification steps. Through these evaluations, it was found that a majority of PLOVER's 16 `event type` categories obtained accuracy rates in the 80%-90% range. This is roughly 10 percent better than ACCENT did against human coding of distinct CAMEO event categories. With regards to entity recovery, a series of separate evaluations of NGEC's fine-tuned RoBERTa QA model performance was performed using a sample of newly collected gold standard-labels for POLECAT's `actor`, `recipient`, date, and location components. RoBERTa-QA was found to obtain accuracy levels of 89.27, 68.64, 71.19, and 69.49 in these cases.[16] Together these evaluations suggest sufficient accuracy for the deployment of NGEC to code POLECAT's full 2010-2022 news corpus.

The full coding pipeline for POLECAT proceeded as follows. First, all politically relevant

[14] Gold standard comparisons suggested agreement with human coders in over 95% of all 952 cases evaluated, though this figure is likely inflated by the rarity of many `context` and `mode` categories.

[15] Specifically, DistilBERT, BERT, DistilRoBERTa and RoBERTa.

[16] This can be contrasted with the those obtained from other language models such as, e.g., BERT, which obtained comparable accuracy levels of 78.81, 61.86, 68.36, and 69.21.

news stories were retrieved from Factiva and translated to English. For entity, `modes`, and `event types`, NGEC then preprocessed stories to remove datelines and embedded newslines, and to reduce the portion of coded text to correspond to the remaining first 512 tokens in each story after these preprocessing steps were applied. Subsequent steps then extract relevant parts-of-speech (POS), datelines, and entity names (i.e., persons, organizations, and locations).[17] Following this, the core distilBERT, SVM, and RoBERTa-QA models are implemented when treating all human annotations as training data. Here, distilBERT and SVMs are first used extract any actual `event types` and `contexts` from a particular news story, respectively. For each extracted `event type`, SVMs are used to recover an event's `modes`; and RoBERTa-QA is used to recover that event's candidate location(s), date(s), `actor(s)` and `recipient(s)`. An offline version of Wikipedia is leveraged for `actor` and `recipient` resolution alongside an expanded version of CAMEO's agents file for more general references to actors such as "soldiers." This ensures that each `actor` and `recipient` is resolved to its proper top-level country code and sector code(s). Event geolocation resolution leverages Mordacai v3 (Halterman, 2017) with the aid of GeoNames (GeoNames, 2022). Alongside each resolved event geolocation (i.e., in terms of latitude-longitude coordinates) a host of additional geolocation information is returned in relation to the raw name of the location reference in the text, location's associated geopolitical unit(s), the location's level of geolocation accuracy, and GeoNames metadata.

Final Pipeline

The above steps generate all relevant PLOVER variable inputs in terms of an event's `event type`, date, location, `actor`, `recipient` and (if applicable) `modes` and `contexts`. One final function is then run over all output to clean the POLECAT data, to add additional variables, and to standardize all relevant keys and value names. In addition to the variables mentioned above, Correlates of War (COW) country codes (COW, 2017) are added, each `actor` and `recipient`'s three-letter country code is (in some versions) included in place of its

[17]These items are used to augment NGEC's primary coding and entity resolution components.

full country name, a PLOVER score-based event intensity measure is calculated and added, a character-string-based quad code is added for each event, and each event's `actors` and `recipients` are converted into lists. The latter information is supplied in several formats, with a JSON format as the primary working version, including primary sector codes for each `actor` and `recipient`, a full list of each `actor` and `recipient`'s sector codes, the raw names of each `actor` and `recipient` as found in the news story's raw text, and (where applicable) stable Wikipedia identifiers for each `actor` and `recipient`. Additional variables are then added for all organizations, individuals, and locations identified within a news story text based upon the POS step described above, as well as for the attributes associated with a coded event's overarching news story. In the latter case, POLECAT supplies details on the associated news story's headline, source, publication date, original language, and identifier. A full list of POLECAT's variables—with definitions—appears in the Supplemental Appendix.

Comparing POLECAT to ICEWS

We compare a preliminary release of POLECAT to ICEWS for the period June 1st, 2020 - February 13th, 2023. While ICEWS itself cannot be seen as "ground truth," past users of ICEWS will be interested in verifying the performance of POLECAT to the existing ICEWS standard. In addition, because the source texts for ICEWS and POLECAT are effectively identical for this recent coding period, comparisons between ICEWS and POLECAT allow us to evaluate differences in the performance of each underlying coder, while holding news content constant. These observations motivate the present comparisons.

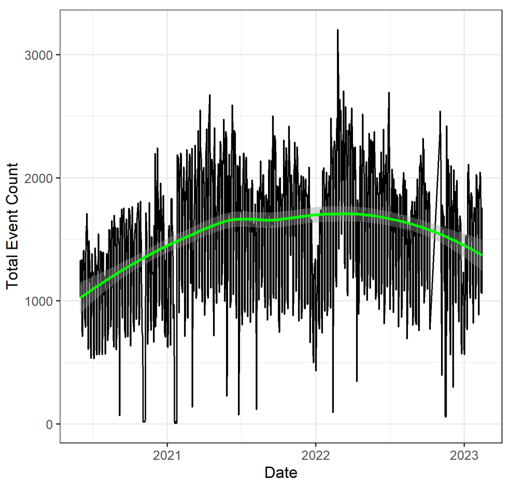
We start with a comparison of ICEWS and POLECAT for the overall events coded within each respective dataset during the June 1st, 2020 - February 13th, 2023 period. Before doing so, we process each dataset in several manners. In an effort to provide end users with the most disaggregated data possible, POLECAT's raw event record output records individual entries for each `event type-mode` combination. This means that specific events of a certain `event type` that have multiple identified modes (e.g., a PROTEST event exhibiting both a demonstration mode and an obstruction mode) are recorded separately in POLECAT. To ensure comparability to ICEWS, we collapse such POLECAT instances into single event records before implementing our comparisons. That being said, news stories that yield multiple events of different `event types`, or multiple events of the same `event type-mode` combinations but involving different `actor(s)` and `recipient(s)` are kept separate. Next, we perform a mild version of "one-a-day" deduplication for both ICEWS and POLECAT, such that the same `actor-recipient`-location triplet can only exhibit a maximum of one event of a certain `event type`(`-mode`-combination) on a particular calendar day.

For all remaining POLECAT and ICEWS events, we first examine each dataset's total events (i) in the aggregate and broken down by quad categories (Table 2) and (ii) over time (Figure 1). Turning to these outputs, we can first observe in Figure 1 and the top portion of Table 2 that POLECAT records several times more total events than ICEWS. This appears to be especially pronounced for the second half of 2020 (Figure 1), and for material cooperation events (Table 2). One contributing factor to this divergence lies in ICEWS' tendency to not record events that could not be geolocated to at least the country-level. POLECAT does not impose this restriction, and retains a large number of events that do not have a location. Two factors likely explain POLECAT's (and other event coders') inability to assign geolocations. First, some event types—such as certain international cooperative interactions—do not always have an inherent location of occurrence. Second, unlike ICEWS, POLECAT refrains from geolocating events based upon the filing city from a news story's byline, so as to avoid false positive geolocations to (e.g.) country capitals. These points notwithstanding, as the final two rows of Table 2 demonstrate, our event totals for POLECAT and ICEWS become much more comparable if one restricts events to only those geolocated to at least the country level of geographic specificity.

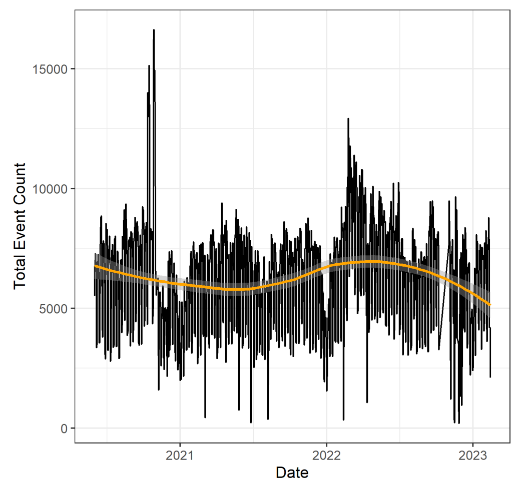
Next, Table 3 shows the distribution of CAMEO and PLOVER `event types` for the overall time period outlined above. These `event types` are not all directly comparable since the underlying definitions can be different, but there are two general patterns to highlight. First,

	Total Total	Verbal Cooperation	Material Cooperation	Verbal Conflict	Material Conflict
ICEWS	1,519,464	613,431	42,559	203,849	259,975
POLECAT	6,223,299	1,440,380	1,114,429	2,040,985	1,627,505
ICEWS Geolocated	1,519,291	613,333	42,556	203,836	259,973
POLECAT Geolocated	2,141,933	487,144	310,123	793,965	535,745

Table 2: Total Event Comparisons, ICEWS and POLECAT, 06/01/2020-02/13/2023



(a) Daily ICEWS Events



(b) Daily POLECAT Events

Figure 1: Global Daily Events Across ICEWS and POLECAT, 06/01/2020-02/13/2023

a quarter of the events in the ICEWS data are CONSULT events, whereas in POLECAT they are under a more moderate tenth. This is likely because of how the two systems handle multi-participant events. While in ICEWS such events are split out into a larger number of pair-wise events, POLECAT and PLOVER handle them organically by allowing multiple entities as the actors or recipients in an event. For example, while a 4-state multilateral meeting would result in up to twelve ICEWS directed-dyad events, in POLECAT it would be retained as a single event with four participants. The second pattern in the event type distributions is that events are more evenly distributed in POLECAT than they are in ICEWS. CONSULT and DEMAND events make up a full 40% of ICEWS, and there are very few material cooperation events (e.g., PROVIDE AID). In POLECAT the distribution is not as skewed. We can most starkly see this by revisiting Table 2 and comparing the material

ICEWS		POLECAT	
CAMEO Event Type	%	PLOVER Event Type	%
CONSULT	24.4	REQUEST	10.3
DEMAND	16.5	ACCUSE	9.2
EXPRESS INTENT TO COOPERATE	8.8	THREATEN	9.1
ENGAGE IN DIPLOMATIC COOPERATION	8.5	CONSULT	8.3
COERCE	7.0	RETREAT	6.2
DISAPPROVE	6.9	AID	6.1
FIGHT	5.3	COERCE	5.9
USE UNCONVENTIONAL MASS VIOLENCE	5.1	PROTEST	5.7
PROTEST	3.3	MOBILIZE	5.5
ASSAULT	2.8	SUPPORT	5.5
REJECT	2.1	ASSAULT	5.5
REDUCE RELATIONS	1.5	REJECT	5.4
THREATEN	1.4	COOPERATE	4.7
APPEAL	1.2	SANCTION	4.7
PROVIDE AID	1.1	AGREE	4.6
YIELD	1.0	CONCEDE	3.3
INVESTIGATE	1.0		
MAKE PUBLIC STATEMENT	0.9		
EXHIBIT FORCE POSTURE	0.5		
ENGAGE IN MATERIAL COOPERATION	0.5		

Table 3: Distribution of events by event type, 06/01/2020-02/13/2023

cooperation rates, which are very rare in ICEWS but quite common in POLECAT. This reflects changes in both PLOVER and NGEC to redefine POLECAT's event type categories and how they are coded.

Regarding geographic location, of those events that are geolocated, both POLECAT and ICEWS are highly concentrated in the sense that most events take place in a small number of countries.[18] Their Simpson index (Simpson, 1949) values are 2.5% and 3.1% respectively. This corresponds to the probability that any two randomly picked events will have been from the same country. POLECAT is slightly more evenly distributed than ICEWS. This is also true at the subnational level if we look at the specific location coded in each event dataset, which is interesting because POLECAT has slightly fewer distinct locations than

[18]Or subnational locations, if we look below the country-level.

Country	ICEWS (%)	POLECAT (%)
India	10.5	3.3
Ukraine	4.9	8.5
United States	4.9	7.8
Russian Federation	7.9	4.5
China	4.7	2.4
Iran	2.1	3.1
United Kingdom	2.9	1.4
Syria	0.6	3.2

Table 4: Countries with more than 2.5% of respective events in either ICEWS or PLOVER, 06/01/2020-02/13/2023

ICEWS (26 thousand versus 33 thousand) for the period considered here, despite having about a third more events in total. This is probably because, as mentioned further above, in POLECAT events without a clearly identifiable location are not coded as having occurred in the story byline's city.

There are more noticeable differences between the countries ICEWS and POLECAT "pay attention to". Table 4 shows the eight countries with more than 2.5% of a given dataset's events, for both ICEWS or POLECAT. ICEWS has relatively more events in India,[19] Russia, China, and the UK, while POLECAT has more in Ukraine, the US[20], Iran, and Syria.

Lastly, Table 5 shows the distribution of contexts in the POLECAT data. The story-level contexts have no equivalent in the ICEWS data, but allow filtering of events in POLECAT based on specific topic areas. As noted earlier, recall that `contexts` are not mutually exclusive in the sense that a story and its associated events can have one, more than one, or no `context` label(s). As can be seen in Table 5, approximately 40% of POLECAT events have no identified `context`, whereas the remaining 60% received at least one `context` tag.

[19]The continued disproportional presence of India in ICEWS is probably due to the legacy bias of the DARPA-ICEWS actor dictionaries, even after the search terms have been modified to put less emphasis on Asia.

[20]US events only include those involving an international, non-US actor, not domestic events.

Events with no context: 40.8%
At least one context: 59.2%, of which:

Context	%	Context	%
diplomatic	13.1	reparations	4.2
rights_freedoms	11.6	pro_autocracy	4.0
territory	11.0	election	4.0
human_rights	8.3	legislative	3.4
pro_democracy	7.9	gender	3.3
political_institutions	7.6	environment	2.7
military	7.5	human_security	2.1
natural_resource	7.2	inequality	1.9
health	6.9	peacekeeping	1.9
legal	6.7	corruption	1.7
technology	6.6	crime	1.5
economic	6.0	lgbt	1.3
terrorism	5.9	misinformation	0.9
migration	5.4	cyber	0.8
intelligence	5.1	asylum	0.6
repression	5.0	illegal_drugs	0.5
religion_ethnicity	4.3	disasters	0.2

Table 5: Distribution of context labels in PLOVER, 06/01/2020-02/13/2023

Case Study: Russia-Ukraine

We now turn to a pair of more focused comparisons of POLECAT to ICEWS. Together these comparisons allow us to compare each event dataset under conditions that better reflect how end users are likely to use POLECAT. For the first such comparison, we retain only material conflict events initiated by Russia-specific `actors` against Ukraine-specific `recipients` during the June 1st 2020 to February 13th 2023 period. As the latter half of this period encompasses Russia's invasion of Ukraine, this comparison will facilitate evaluations of how well each event dataset captures this widely understood increase in material conflict between Russia and Ukraine. Figure 2 presents several relevant plots from the corresponding Russia $\rightarrow$ Ukraine material conflict events obtained separately from POLECAT and ICEWS. Subfigures 2a-2b plot the daily event counts for this directed material conflicts dataset over time. Following a period of relatively little material conflict, each of these subfigures indicates a significant spike in material conflict coinciding with Russia's February 2022 invasion of Ukraine, followed by a sustained, albeit declining, level of material conflict intensity thereafter. Both time series are remarkably similar, a quality that is further verified by their plots against one another in Subfigure 2c.[21] Finally, we map the corresponding Russia-directed material conflict events for those events that had city-level geolocation precision in Subfigures 2d-2e. Here we again find comparable results for ICEWS and POLECAT, albeit fewer geolocations for POLECAT overall, owing at least in part to its more conservative geolocation routines. In sum, an assessment of international material conflict events initiated by Russia against Ukraine suggests that POLECAT is highly similar to ICEWS in recovered events, albeit with fewer geolocated events overall.

Case Study: Iran Protests

Our second case study compares POLECAT and ICEWS for a specific form of intra-state, rather than inter-state, conflict. In this case, we focus on Iran-based protests for the June 1st, 2020 to February 13th, 2023 period. Iran experienced a number of significant protests

[21] The points on this latter subfigure, and the comparable subfigure in Figure 3 below, were jittered.

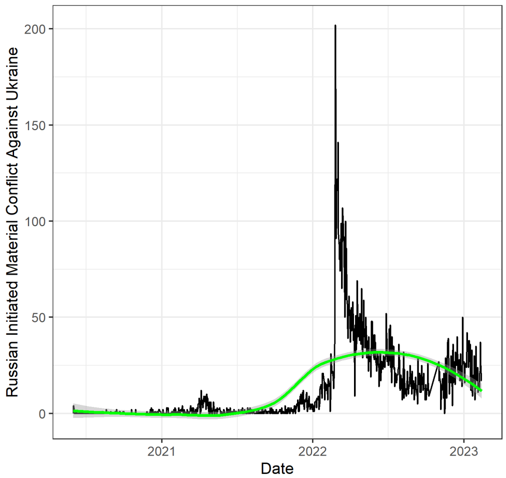
during this period, most notably (i) protests in the latter half of 2020 related to food- and economic-concerns, (ii) an intensification of these protests during the January-March 2021 period led in part by retirees and pensioners, (iii) water- and economic-related protests during the latter half of 2021, (iv) food protests beginning in May 2022, and (v) women's rights protests beginning in Fall 2022.

For our comparisons, we retain only ICEWS and POLECAT events that were recorded as arising from their respective `PROTEST event type` categories, that were initiated by Iranian `actors`, and that were geolocated to Iran or a subnational unit therein. We provide a series of plots comparing the corresponding events in Figure 3. Starting with our daily time series counts of Iranian protests for ICEWS and POLECAT in Subfigures 3a-3c, we find less correspondence between these event datasets than was the case for our comparison of Russia-Ukraine material conflict. One key distinction appears to be POLECAT's recording of heightened protest activity in early 2021 that ICEWS did not capture. This period corresponds to a spike in violent protests in Iran associated with deteriorating economic conditions, including a January 24th stock market crises and related pressures on pension funds, fuel prices, and other goods.[22] In February 2021 alone, 271 protests were reported in 67 cities across Iran (Hamidi, 2021; HRW, 2021; INW, 2021). These protests then continued into March before dying down in April 2021. That period aside, both datasets imply an upward trend in Iranian protest activity during the June 1st, 2020 to February 13th, 2023 period, with a spike in protest activity in relation to Iran's women's rights protests in late 2022. Geographic maps of city-level geolocated protest events in Subfigures 3d-3e likewise suggest a degree of correspondence between our ICEWS and POLECAT protest events, albeit again with more geographic dispersion in event geolocations for ICEWS. Altogether, these results again suggest a number of similarities between POLECAT and ICEWS for studying protest events, with the following two caveats. First, POLECAT appears to have captured at least some relevant instances of Iranian protest that ICEWS missed. Second, ICEWS continues to more extensively record city-level geolocated events than does POLECAT, albeit

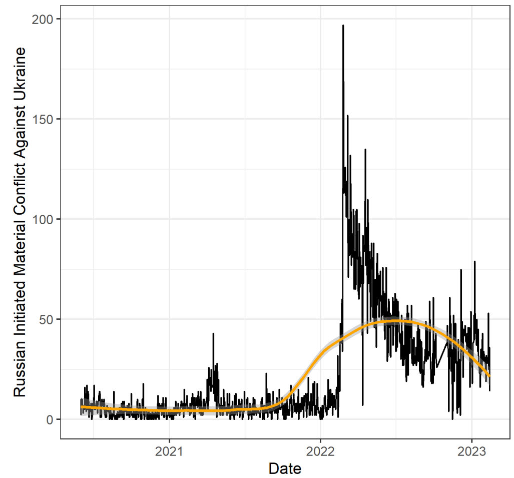
[22]See, e.g., Fazeli (2021) and the sources cited below.

with an added potential for false positives in these geolocations.

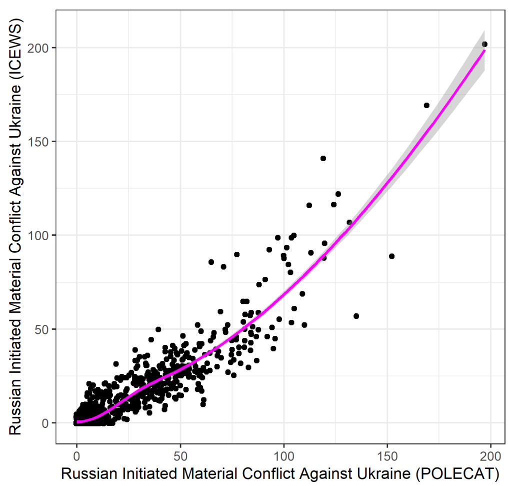
In sum, an initial test release of POLECAT performs comparably to ICEWS along several classic event data dimensions, both in the aggregate and for a number of more focused comparisons related to both interstate and intrastate political conflict. For example, the comparison of ICEWS to POLECAT for the Russia-Ukraine conflict suggested that both datasets are remarkably similar in the events they capture. At the same time, we found that POLECAT also offers additional unique information in several respects. This is most notable in POLECAT's `context` categories, which are not available in ICEWS nor in many other widely used event datasets. Alongside this, we also found that POLECAT identified at least some relevant some protest instances in our Iran case study that ICEWS did not. This suggests that POLECAT has the potential to recover more accurate records of political events from comparable news coprora to ICEWS, though as noted earlier the geolocation of these events remains more conservative in POLECAT at present.



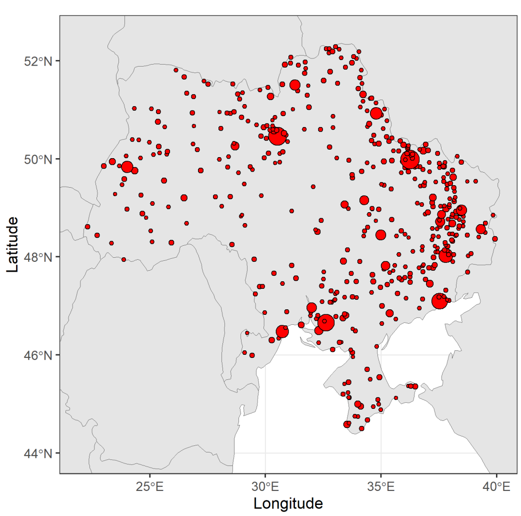
(a) Daily ICEWS Events



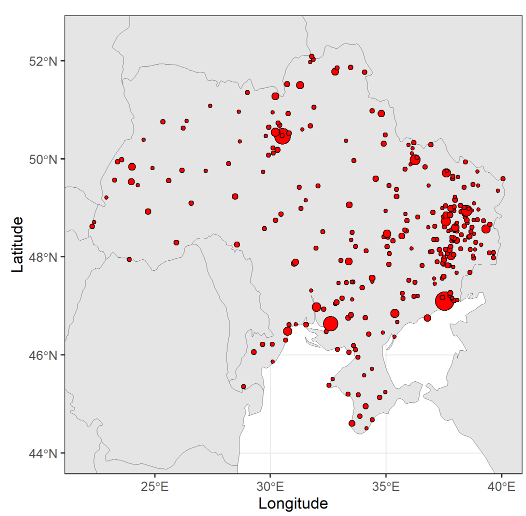
(b) Daily POLECAT Events



(c) POLECAT vs. ICEWS Daily Events

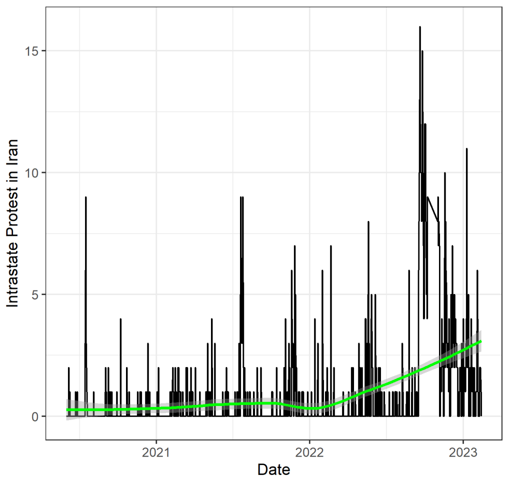


(d) Geolocated ICEWS Events

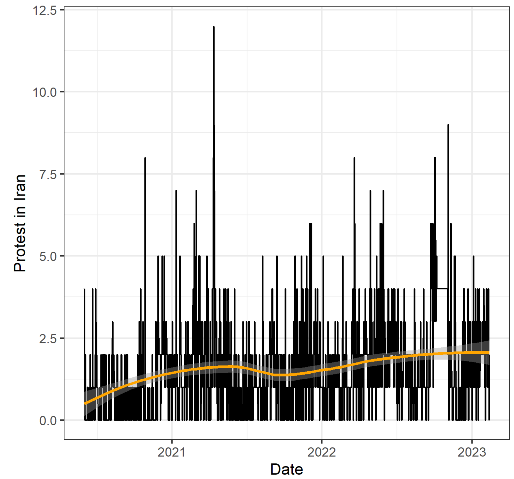


(e) Geolocated POLECAT Events

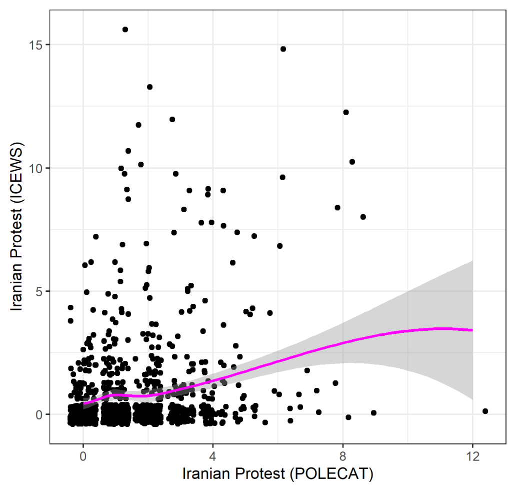
Figure 2: Russia->Ukraine Material Conflict Events Across ICEWS and POLECAT, 06/01/2020-02/13/2023



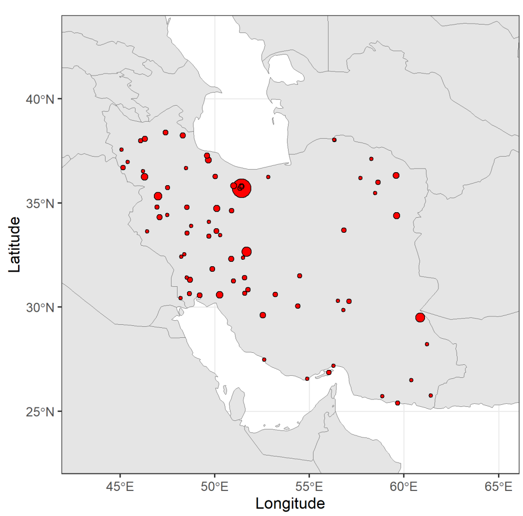
(a) Daily ICEWS Events



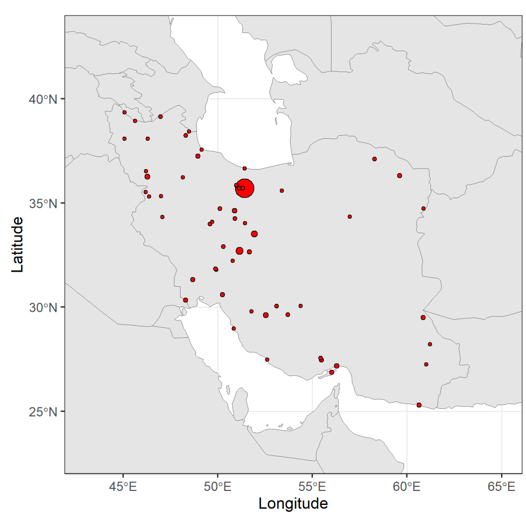
(b) Daily POLECAT Events



(c) POLECAT vs. ICEWS Daily Events



(d) Geolocated ICEWS Events



(e) Geolocated POLECAT Events

Figure 3: Iranian Protest Events Across ICEWS and POLECAT, 06/01/2020-02/13/2023

Conclusion

Existing automated event data systems have proven difficult to maintain and update. This represents a significant challenge for automated event data coding in light of (i) the changing nature of global social and political events and (ii) recent innovations in automated text extraction methodologies. This paper addresses these challenges by first presenting a new, flexible event data ontology known as PLOVER, which simplifies previous event ontologies' event types while providing unique capabilities for coding additional information on the "how" and "why" of each event. Together these refinements foster both more accurate and more contextualized event records. We illustrate PLOVER's potential by using it within the next generation event coder (NGEC) to generate a new global event dataset known as POLECAT. POLECAT is a fully automated event dataset that is intended to succeed the ICEWS event dataset. In contrast to extant automated event coding systems, the coder used to generate POLECAT does not rely on actor and event dictionaries. Rather, the coder modularly employs easy to update machine learning methods.

Future Extensions

The above points notwithstanding, the present analysis represents only the first step in developing PLOVER and POLECAT. Future iterations of this work will more extensively validate POLECAT against additional datasets and data sources, drawing upon past research designs for internally and externally validating political event data (e.g., Ward et al., 2013; Bagozzi et al., 2019; Stundal et al., 2021; Althaus, Peyton and Shalmon, 2022). After the data coding pipeline for POLECAT has stabilized, we also anticipate releasing benchmark annotated text data. The underlying text for these benchmark data will be synthetic—that is, cases almost indistinguishable from actual new stories but artificially generated using the transformer models, a task at which they excel—in order to avoid intellectual property issues. Though challenging to create, benchmark datasets are useful in spurring additional research and innovation in the area of automated event coding. In event extraction, one of the standard datasets in natural language processing (NLP) is the DARPA-funded ACE dataset, which has been used in more than a thousand publications. Making a large set of high-quality annotated data available will encourage researchers to continue to innovate on the PLOVER, POLECAT, and NGEC innovations presented above.

Closely related to the idea of benchmark datasets is the notion of further improvements and refinements to the annotated stories that are used as training inputs for POLECAT. Regarding human annotation, future anticipated refinements will include (i) improvements to current annotations[23] with the aid of automated detection methods and (ii) additional annotations for especially challenging `event types`, `modes`, and `contexts`, as well as more annotations for actor, recipient, location, and date spans. The additions and refinements will first be made to address any deficiencies in the initial wave of annotations after the full POLECAT pipeline is stabilized. In the longer term, we also anticipate that future waves of expert annotation will be necessary, especially based upon end-user feedback and the changing nature of international `events`, `modes`, and `contexts`.

We also plan future extensions with regard to hyper-parameter optimization. Hyper-parameters are internal parameters used in the SVM, transformer-based event classifier, and QA modeling processes that cannot be estimated from the data. The initial POLECAT release that we analyze above is primarily using the default values for these hyper-parameters across the various machine learning models that NGEC leverages. Better hyper-parameter selection will likely increase the accuracy of each stage of the NGEC machine classification pipeline and can be implemented through a variety of out-of-sample comparisons of current model performance with some additional computation. We also plan to look both at existing alternatives to distilBERT and RoBERTa and at subsequent models that are likely to emerge into the methodological foreground for similar text-as-data tasks. Given the tremendous recent interest generated by the ChatGPT system that is based on the GPT-3 large language model, and the billions of dollars of new investment going into related models, at least some of which will be open source in whole or in part, we anticipate substantial new opportunities for further enhancing the system with improved software in coming months and years.

[23] E.g., via the identification and removal of some outlier cases or imperfect annotations.

One final extension that we hope to develop corresponds to the addition of `modes` and/or `contexts` for negations, hypotheticals, future tenses, and claims. These event modalities are ones that current event coders do not do a good job of picking up. In most cases, this is because the modalities mentioned largely convey uncertainty about whether the event actually happened. Additional `modes` and/or `contexts` for negations, hypotheticals, future tense, and claims would address this issue. These could be identified with human coding and/or a machine learning classifier, though doing this will likely require more data annotation efforts.

References

Althaus, Scott, Buddy Peyton and Dan Shalmon. 2022. "A Total Error Approach for Validating Event Data." *American Behavioral Scientist* 66(5):603–624.

Althaus, Scott, Joseph Bajjalieh, John F. Carter, Buddy Peyton and Dan A. Shalmon. 2019. "Cline Center Historical Phoenix Event Data." Cline Center for Advanced Social Research. v.1.3.0. University of Illinois Urbana-Champaign. May 4. `https://doi.org/10.13012/B2IDB-0647142_V3`.

Azar, Edward E. 1980. "The Conflict and Peace Data Bank (COPDAB) Project." *Journal of Conflict Resolution* 24:143–152.

Azar, Edward E. 1982. *The Codebook of the Conflict and Peace Data Bank (COPDAB)*. College Park, MD: Center for International Development, University of Maryland.

Azar, Edward E. 2009. "Conflict and Peace Data Bank (COPDAB), 1948-1978." Inter-university Consortium for Political and Social Research [distributor], 2009-06-03. https://doi.org/10.3886/ICPSR07767.v4.

Azar, Edward E. and Thomas Sloan. 1975. *Dimensions of Interaction*. Pittsburgh: University Center for International Studies, University of Pittsburgh.

Bagozzi, Benjamin E., Patrick T. Brandt, John R. Freeman, Jennifer S. Holmes, Alisha Kim, Agustin Palao Mendizabal and Carly Potz-Nielsen. 2019. "The Prevalence and Severity of Underreporting Bias in Machine- and Human-Coded Data." *Political Science Research and Methods* 7(3):641–649.

Beger, Andreas, Richard K. Morgan and Michael D. Ward. 2021. "Reassessing the Role of Theory and Machine Learning in Forecasting Civil Conflict." *Journal of Conflict Resolution* 65(7-8):1405–1426.

Beieler, John, Patrick T Brandt, Andrew Halterman, Erin Simpson and Philip A Schrodt. 2016. Generating Political Event Data in Near Real Time: Opportunities and Challenges.

In *Computational Social Science*, ed. R. Michael Alvarez. Cambridge University Press chapter 98.

Blair, Robert A. and Nicholas Sambanis. 2020. "Forecasting Civil Wars: Theory and Structure in an Age of "Big Data" and Machine Learning." *Journal of Conflict Resolution* 64(10):1885–1915.

Bond, Doug, J. Craig Jenkins, Charles L. Taylor Taylor and Kurt Schock. 1997. "Mapping Mass Political Conflict and Civil Society: Issues and Prospects for the Automated Development of Event Data." *Journal of Conflict Resolution* 41(4):553–579.

Bond, Doug, Joe Bond, Churl Oh, J. Craig Jenkins and Charles Lewis Taylor. 2003. "Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development." *Journal of Peace Research* 40(6):733–745.

Boschee, Elizabeth, Jennifer Lautenschalger, Sean O'Brien, Steve Shellman, James Starz and Michael Ward. 2015. "ICEWS Coded Event Data." `https://doi.org/10.7910/DVN/28075`, Harvard Dataverse, V35.

Chiba, Daina and Kristian Skrede Gleditsch. 2017. "The shape of things to come? Expanding the inequality and grievance model for civil war forecasts with event data." *Journal of Peace Research* 54(2):275–297.

Colaresi, Michael. 2004. "When Doves Cry: International Rivalry, Unreciprocated Cooperation, and Leadership Turnover." *American Journal of Political Science* 48(3):555–570.

Cortes, Corinna and Vladimir Vapnik. 1995. "Support-vector networks." *Machine learning* 20(3):273–297.

COW. 2017. "State System Membership List, v2016." Correlates of War Project: `http://correlatesofwar.org`.

Devlin, J., Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Du, Xinya and Claire Cardie. 2020. "Event Extraction by Answering (Almost) Natural Questions." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 671–683.

Fazeli, Yaghoub. 2021. "Iran Protesters Storm Iran State Building After Deadly Shootings Near Pakistan Border." Al Arabiya English, February 23rd, 2021. https://english.alarabiya.net/News/middle-east/2021/02/23/Iran-protests-Protesters-storm-Iran-state-building-after-deadly-shootings-near-Pakist

FitzGerald, Nicholas, Julian Michael, Luheng He and Luke Zettlemoyer. 2018. "Large-scale QA-SRL parsing." *arXiv preprint arXiv:1805.05377* .

GeoNames. 2022. "GeoNames." https://www.geonames.org/.

Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco and Judith L. Weddle. 1994. "The Machine Coding of Events from Regional and International Sources." *International Studies Quarterly* 38(1):91–119.

Goldstein, Joshua S. 1992. "A Conflict-Cooperation Scale for WEIS Events Data." *Journal of Conflict Resolution* 36:369–385.

Goldstein, Joshua S and John R Freeman. 1990. *Three Way Street: Strategic Reciprocity in World Politics*. University of Chicago Press.

Grant, Christan, Andrew Halterman, Jill Irvine, Yan Liang and Khaled Jabr. 2019. "TERRIER (Temporally Extended, Regular, Reproducible International Event Records) BETA." OU Event Data Project https://terrierdata.org/.

Halterman, Andrew. 2017. "Mordecai: Full text geoparsing and event geocoding." *Journal of Open-Source Software* 2(9):91.

Halterman, Andrew. 2022. "Synthetically generated text for supervised text analysis." *PolMeth conference paper* .

Halterman, Andrew, Jill Irvine, Christan Grant, Khaled Jabr and Yan Liang. 2018. "Creating an Automated Event Data System for Arabic Text." Prepared for Presentation at the ISA Annual Meeting 2018 in San Francisco.

Halterman, Andrew, Philip A. Schrodt, Andreas Beger, Benjamin E. Bagozzi and Grace I. Scarborough. 2023. "Creating Custom Event Data Without Dictionaries: A Bag-of-Tricks." Working Paper.

Hamidi, Mahmoud. 2021. "Iranians Continue Protests; at Least Six Rallies and Strikes on February 9." Iranian News Update, March 10th, 2021. https://irannewsupdate.com/news/insider/iranians-continue-protests-at-least-six-rallies-and-strikes-on-february-9/?amp=.

HRW. 2021. "Iran: Killings Near Pakistani Border Alleged Excessive Use of Force, Internet Shutdown." Human Rights Watch, February 25th, 2021. https://www.hrw.org/news/2021/02/25/iran-killings-near-pakistani-border.

INW. 2021. "February Iran Protests Report: 271 protests in 67 cities." Iran News Wire (INW), March 6th, 2021. https://irannewswire.org/february-iran-protests-report-271-protests-in-67-cities/.

Kibris, Arzu. 2021. "The geo-temporal evolution of violence in civil conflicts: A micro analysis of conflict diffusion on a new event data set." *Journal of Peace Research* 58(5):885–899.

King, Gary and Will Lowe. 2003. "An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57(3):617–642.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omar Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pre-Training Approach." arXiv:1907.11692. arXiv preprint.

McClelland, Charles. 2006. "World Event/Interaction Survey (WEIS) Project, 1966-1978." Inter-university Consortium for Political and Social Research [distributor], 2006-01-18. https://doi.org/10.3886/ICPSR05211.v3.

McClelland, Charles A. 1967. "World-Event-Interaction-Survey: A Research Project on the Theory and Measurement of International Interaction and Transaction." University of Southern California.

McClelland, Charles A. 1976. *World Event/Interaction Survey Codebook (ICPSR 5211)*. Ann Arbor: Inter-University Consortium for Political and Social Research.

Norris, Clayton, Philip Schrodt and John Beieler. 2017. "PETRARCH2: Another Event Coding Program." *The Journal of Open Source Software* 2.

O'Brien, Sean P. 2010. "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research." *International Studies Review* 12(1):87–104.

Open Event Data Alliance. 2015. "Phoenix Near-Real-Time Data." `http://openeventdata.org/datasets.html`.

Osorio, Javier, Viveca Pavon, Sayeed Salam, Jennifer Holmes, Patrick T. Brandt and Latifur Khan. 2019. "Translating CAMEO Verbs for Automated Coding of Event Data." *International Interactions* 45(6):1049–1064.

Radford, Benjamin J. 2021. "Automated Dictionary Generation for Political Eventcoding." *Political Science Research and Methods* 9:157–171.

Reuveny, Rafael and Heejoon Kang. 1996. "International Trade, Political Conflict/Cooperation, and Granger Causality." *American Journal of Political Science* 40(3):943–970.

Sanh, Victor, Lysandre Debut, Julien Chaumond and Thomas Wolf. 2019. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.".
URL: *https://arxiv.org/abs/1910.01108*

Schrodt, Philip A. 2012. "CAMEO: Conflict and Mediation Event Observations Event and Actor Codebook." Manual. Version: 1.1b3, March 2012, https://parusanalytics.com/eventdata/cameo.dir/CAMEO.Manual.1.1b3.pdf.

Schrodt, Philip A., Deborah J. Gerner and Ömür Yilmaz. 2009. Conflict and Mediation Event Observations (CAMEO): An Event Data Framework for a Post Cold War World. In *International Conflict Mediation: New Approaches and Findings*, ed. Jacob Bercovitch and Scott Gartner. New York: Routledge.

Schrodt, Philp A. 2011. "TABARI: Textual Analysis by Augmented Replacement Instructions." Manual. `https://parusanalytics.com/eventdata/tabari.dir/tabari.manual.0.7.6.pdf`, Version 0.7.6.

Simpson, E.H. 1949. "Measurement of Diversity." *Nature* 163:688.

Stundal, Logan, Benjamin E. Bagozzi, John R. Freeman and Jennifer S. Holmes. 2021. "Human Rights Violations in Space: Assessing the External Validity of Machine-Geocoded versus Human-Geocoded Data." *Political Analysis* p. 1–17.

Ward, Michael D., Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff and Ben Radford. 2013. "Comparing GDELT and ICEWS event data." Working Paper.

Weschle, Simon. 2018. "Quantifying Political Relationships." *American Political Science Review* 112(4):1090–1095.

Windsor, Leah C. 2022. "Bias in Text Analysis for International Relations Research." *Global Studies Quarterly* 2:1–13.