

An Ordinal Latent Variable Model of Conflict Intensity

Niklas Stoehr[‡] Lucas Torroba Hennigen[‡] Josef Valvoda[‡]

Robert West[‡] Ryan Cotterell[‡] Aaron Schein[‡]


[‡]ETH Zürich [‡]MIT [‡]University of Cambridge [‡]EPFL [‡]The University of Chicago

niklas.stoehr@inf.ethz.ch lucastor@mit.edu jv406@cam.ac.uk

robert.west@epfl.ch ryan.cotterell@inf.ethz.ch schein@uchicago.edu

Abstract

Measuring the intensity of events is crucial for monitoring and tracking armed conflict. Advances in automated event extraction have yielded massive data sets of “who did what to whom” micro-records that enable data-driven approaches to monitoring conflict. The Goldstein scale is a widely-used expert-based measure that scores events on a conflictual-cooperative scale. It is based only on the action category (“what”) and disregards the subject (“who”) and object (“to whom”) of an event, as well as contextual information, like associated casualty count, that should contribute to the perception of an event’s “intensity”. To address these shortcomings, we take a latent variable-based approach to measuring conflict intensity. We introduce a probabilistic generative model that assumes each observed event is associated with a latent intensity class. A novel aspect of this model is that it imposes an ordering on the classes, such that higher-valued classes denote higher levels of intensity. The ordinal nature of the latent variable is induced from naturally ordered aspects of the data (e.g., casualty counts) where higher values naturally indicate higher intensity. We evaluate the proposed model both intrinsically and extrinsically, showing that it obtains good held-out predictive performance.

 <https://github.com/niklasstoehr/ordinal-conflict-intensity>

1 Introduction

On a scale from -10 for conflictual to $+10$ for cooperative, which of the following events should be considered more “intense”: “Soldiers injured two civilians” or “Rebels detained fifty soldiers”?

Measuring the intensity of events is crucial for monitoring and tracking armed conflict. Advances in the automated collection and coding of events have produced massive and systematized data sets of micro-records that enable data-driven

CAMEO code	action name	Goldstein value	avg. # casualties
19	fight	-10.0	9.31
20	mass violence	-10.0	42.20
18	assault	-9.0	11.47
15	force posture	-7.2	0.13
17	coerce	-7.0	1.44
14	protest	-6.5	2.06
13	threaten	-6.0	0.13
10	demand	-5.0	0.01
12	reject	-4.0	0.00
16	reduce relations	-4.0	0.00
9	investigate	-2.0	0.04
11	disapprove	-2.0	0.03
1	public statement	0.0	0.19
4	consult	1.0	0.03
2	appeal	3.0	0.00
5	diplom cooperation	3.5	0.03
3	intent cooperate	4.0	0.05
8	yield	5.0	0.10
6	material cooperation	6.0	0.01
7	provide aid	7.0	0.00

Table 1: The **The Goldstein scale** is an expert-based intensity ranking of the 20 **CAMEO action categories** ranging between -10.0 (conflictual) to $+10.0$ (cooperative). The scale disregards casualty counts that are typically considered in conflict assessment. Here, we display casualty counts as reported in the **NAVCO dataset**.

approaches to monitoring conflict. While the “intensity” of a given event has traditionally been assessed by human expert raters, the tremendous quantity of events collected every day makes case-by-case analysis unmanageable. As a consequence, there is a strong demand for automated and model-based methods to aggregate events into meaningful “conflict intensity” measures.

One of the most frequently used measures is the Goldstein scale (Goldstein, 1992). Major event datasets like IDEA (Bond et al., 2003), KEDS (Schrodt, 2008), GDELT (Leetaru and Schrodt, 2013), ICEWS (Boschee et al., 2015), Phoenix (Beielser, 2016) and NAVCO (Lewis et al., 2016) all rely on it. The Goldstein scale assigns intensity scores between -10.0 and $+10$

on a conflictual–cooperative scale to the action categories defined by the [Conflict and Mediation Event Observations \(CAMEO\)](#) event coding scheme ([Schrodt, 2012](#)). CAMEO specifies 204 low-level event types which are summarized into 20 high-level action categories. The Goldstein scale ranks “use unconventional mass violence” and “fight” as the most conflictual of the 20 high-level action categories (-10.0) and “provide aid” ($+7.0$) as the most cooperative; see [Tab. 1](#).

Despite its usage, the Goldstein scale has many well-known shortcomings ([King and Lowe, 2003](#); [Schrodt, 2019](#)). In particular, it applies only to action categories, and does not account for any contextual information of a given event, like which actors are involved, or how many fatalities resulted, among other bits of context that should contribute to the perception of an event’s “intensity”.

This paper takes a latent-variable based approach to measuring conflict intensity. We introduce a probabilistic generative model that assumes each observed event n is associated with a latent intensity class z_n . A novel aspect of this model is that it imposes an ordering on the classes, such that higher values of z_n denote higher levels of intensity. The ordinal nature of z_n is induced from naturally ordered aspects of the data (e.g., casualty counts) where higher values naturally indicate higher intensity. The model effectively learns to interpolate the ordered (i.e., cardinal or ordinal) elements of the data while inferring correlation structure with the non-ordered (e.g., categorical) elements of the data (e.g., actor types).

We start with a discussion of the Goldstein scale and introduce a political event dataset annotated with Goldstein values in [§2](#) and [§3](#). Then, we propose our model with an ordinal latent variable in [§4](#). We evaluate the performance of the model intrinsically ([§5](#)) and extrinsically ([§6](#)) and find that it improves over measures based on the original Goldstein scale or heuristics based on the raw data.

2 Limitations of the Goldstein Scale

The Goldstein scale is a widely-used measure of the conflictual versus cooperative nature of interactions between countries ([Goldstein, 1992](#)). The scale was created by a panel of international relations experts who ranked descriptions of interactions. It was initially created to score action categories in the WEIS event ontology ([McClelland, 1984](#)) and was later adapted to CAMEO ([Schrodt, 2012](#)).

The Goldstein scale applies only to the action category of an event (e.g., “fight” or “trade”). Thus, two different “fight” events, which might involve two different pairs of actors, occur at different times, or differ dramatically with respect to the number of associated fatalities, will still be assigned the same Goldstein value. The Goldstein scale is thus a poor measure of a conflict’s perceived “intensity”, as it ignores much of the information that contributes to that perception. Recent work in conflict studies, for instance, operationalizes “intensity” primarily using casualty counts ([Chaudoin et al., 2017](#); [Zhong et al., 2023](#)), which the Goldstein scale ignores entirely.

In [Tab. 1](#), we show the empirical distribution of assigned Goldstein values alongside the empirical distribution of casualty counts in a dataset of conflict events. The Goldstein scale is very coarse-grained; while it ostensibly ranges between -10.0 and $+10.0$, only a small number of discrete values ever occur, with many action categories assigned the same value. For the purpose of measuring conflict intensity, a finer-grained and more contextual scale is desirable.

3 Conflict Event Data

This paper considers the publicly available [Non-violent and Violent Campaigns and Outcomes \(NAVCO\)](#) data collection ([Chenoweth et al., 2018](#)), specifically, the latest release NAVCO 3.0 from November 2019 which comprises $N = 112,089$ events between December 1990 and December 2012. An exemplary event description is “On 19 May 2012, **soldiers** **injured** **two** **civilians** in Afghanistan”. Each part of this description has been parsed by human coders into standardized, structural features. We color-code the features that correspond to the semantic roles [subject](#), [predicate](#), [quantifier](#), [object](#), which are the focus of our modeling approach. Each data point n thus consists of a four-element tuple $\{s_n, p_n, q_n, o_n\}$. We note that events are further coded for their location (in this case, Afghanistan) and time (19 May 2012), among other bits of contextual information. Let us discuss each feature in more detail:

Subject s_n . NAVCO contains columns termed “actor3”, “actor6” and “actor9” which code for the subject (or agent) of a given action. The actor types are defined by the [CAMEO actor codebook](#). We first merge the higher-level categories “actor3” and “actor6”, resulting in 33 different actor types, and

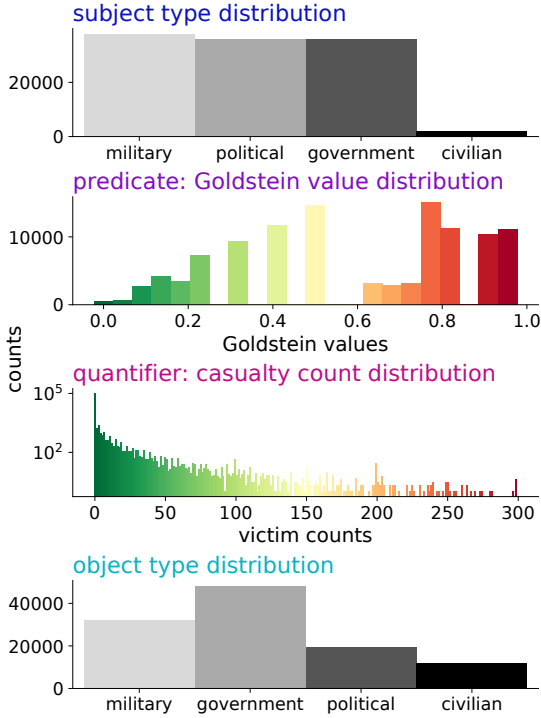


Figure 1: Data distributions of NAVCO 3.0 dataset (Chenoweth et al., 2018). Goldstein values and casualty counts have intrinsic intensity orderings. Goldstein values are reversed and transformed, so that 1.0 represents the most conflictual. We model the subject s_n as Categorical, the predicate p_n as Beta, the quantifier q_n as Zero-inflated Geometric and the object o_n as Categorical.

then map all actor types into one of $S = 4$ classes: $s_n \in \{\text{civilian, military, governmental, political}\}$. We present our 4-class actor type mapping in Tab. 3 of the appendix.

Predicate p_n (Goldstein values). NAVCO codes¹ each event description into one of the 20 CAMEO action categories in the column “verb10”, which is by extension associated with a Goldstein value p_n . Throughout, we refer to the Goldstein value p_n as an action’s “predicate”, since there is a one-to-one mapping between action categories and Goldstein values. We scale Goldstein values p_n to a $[0, 1]$ range and invert them (i.e., $p_n \leftarrow 1 - p_n$) so that higher values close to 1 represent more conflictual action categories.

Quantifier q_n (casualty counts). Each event description is annotated with human-verified fatality and wounded counts. We add the two and

¹NAVCO features a 21st action category which we exclude since it is not specified by the CAMEO taxonomy and thus has no Goldstein value.

refer to the resulting value $q_n \in \mathbb{N}_0^+$ as an event’s “quantifier” or its “casualty count”. In Tab. 1, we give the average number of casualties associated with each action alongside its Goldstein value—as intuition might suggest, actions that Goldstein scores as more conflictual (e.g., “fight” (−10.0)) coincide with more casualties on average.

Object o_n . Similar to its subject, NAVCO codes for the direct object or “target” of a conflict action using the CAMEO coding scheme; these codes are found in the columns “target3” and “target6”. We map those into the $O = 4$ classes so that $o_n \in \{\text{civilian, military, governmental, political}\}$.

Contextual information: location and time. Finally, each event is further annotated with a timestamp and location, which we use to design extrinsic evaluation tasks in §6.

4 Ordinal Latent Variable Model

We operationalize conflict intensity as a latent variable that expresses the association between the observed variables subject (s_n), predicate (p_n), quantifier (q_n) and object (o_n). Each data point is a tuple $\{s_n, p_n, q_n, o_n\}$ representing an event. Our Bayesian latent variable model is depicted in Fig. 2. We assume the following generative story. For each event n , we assume that its event intensity class $z_n \in \{1, \dots, C\}$ is a Categorical random variable

$$z_n \sim \text{Categorical}(\boldsymbol{\pi}^{(z)}) \quad (1)$$

where $\boldsymbol{\pi}^{(z)}$ is a C -dimensional discrete distribution over latent intensity classes. We place a Dirichlet prior over $\boldsymbol{\pi}^{(z)}$

$$\boldsymbol{\pi}^{(z)} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(z)}) \quad (2)$$

with concentration parameter $\boldsymbol{\alpha}^{(z)} \in \mathbb{R}_+^C$. Conditioned on z_n , we assume each of the observed sites per event tuple s_n, p_n, q_n and o_n are then drawn as

$$s_n | z_n \sim \text{Categorical}(\boldsymbol{\pi}_{z_n}^{(s)}) \quad (3)$$

$$p_n | z_n \sim \text{Beta}(\omega_{z_n}^{(p)}, \kappa_{z_n}^{(p)}) \quad (4)$$

$$q_n | z_n \sim \text{Zero-infl. Geom.}(\delta_{z_n}^{(q)}, b_{z_n}^{(q)}) \quad (5)$$

$$o_n | z_n \sim \text{Categorical}(\boldsymbol{\pi}_{z_n}^{(o)}) \quad (6)$$

Here $\boldsymbol{\pi}_{z_n}^{(s)}$ and $\boldsymbol{\pi}_{z_n}^{(o)}$ are the discrete distributions for class z_n over S subject and O object types, respectively. They are given as row vectors in the matrices $\boldsymbol{\Pi}^{(s)} \in (0, 1)^{C \times S}$ and $\boldsymbol{\Pi}^{(o)} \in (0, 1)^{C \times O}$ that z_n

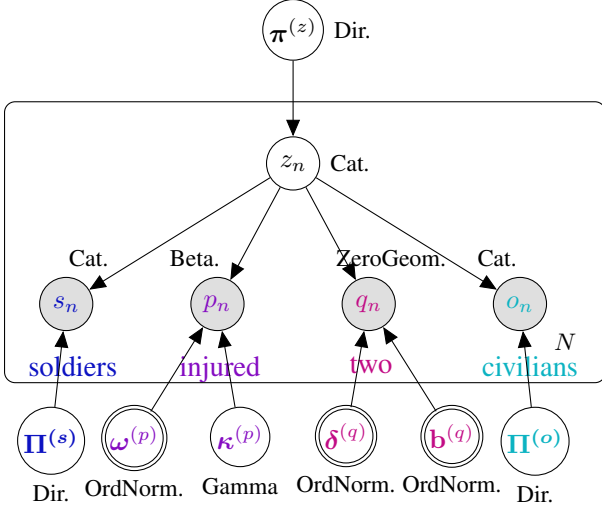


Figure 2: The proposed latent variable model of conflict intensity. The observed sites of an event tuple $\{s_n, p_n, q_n, o_n\}$ describe the **subject type**, **Goldstein value**, **casualty counts**, **object type** in an event description such as “soldiers injured two civilians”. The parameters $\omega^{(p)}$, $\delta^{(q)}$ and $\mathbf{b}^{(q)}$ are ordinally transformed vectors, as indicated by double-border nodes (⊙) which leads the latent z_n to represent ordinal intensity classes.

indexes into. We place a simple symmetric Dirichlet prior over all rows—e.g., $\pi_c^{(s)} \sim \text{Dirichlet}(\mathbf{1}_S)$ for $c \in \{1, \dots, C\}$. The scalar parameters $\omega_{z_n}^{(p)}$, $\kappa_{z_n}^{(p)}$, $\delta_{z_n}^{(q)}$, $\mathbf{b}_{z_n}^{(q)}$ are similarly selected by z_n from C -dimensional vectors $\omega^{(p)}$, $\kappa^{(p)}$, $\delta^{(q)}$, $\mathbf{b}^{(q)}$ as discussed below.

4.1 Ordinal Latent Variable

We want the latent variable z_n to be ordinal, such that higher-valued classes correspond to higher intensity levels. However, in the model thus described, z_n is a Categorical random variable whose classes are not inherently ordered. So how does this model encode an ordinal z_n ? Unlike the subject s_n and object o_n , which are categorical, the Goldstein value p_n and casualty count q_n are cardinal quantities whose magnitudes naturally indicate the “intensity” of a given event. To capture this intuition, we first assume that p_n and q_n are drawn from Beta (eq. (4)) and Zero-inflated Geometric distributions (eq. (5)), respectively, whose parameters are indexed by z_n . We then impose an ordering on the parameters, such that higher classes (e.g., $z_n = c$) correspond to higher-valued parameters (e.g., $\omega_c^{(p)} > \omega_{c-1}^{(p)}$) which in turn encourage larger values of the observed cardinal quantities (e.g., p_n).²

²We note that one might also impose ordering on casualty types (e.g., an event with civilian casualties might be consid-

Ordered Normal To flexibly impose ordering on vectors of parameters, we first define the Ordered Normal prior (Stoehr et al., 2023a). An ordered vector $\lambda = (\lambda_1, \dots, \lambda_C)$, where $\lambda_c > \lambda_{c-1}$, is a C -dimensional Ordered Normal random variable $\lambda \sim \text{OrderedNormal}_C(\mu, \sigma)$ if it is sampled according to the following generative process:

$$x_c \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma) \quad \text{for } c \in \{1, \dots, C\}$$

$$(\lambda_1, \dots, \lambda_C) \leftarrow \text{Ord}(\{x_1, \dots, x_C\}) \quad (7)$$

where $\text{Ord}(\cdot)$ takes in an unordered set of numbers $\{x_1, \dots, x_C\}$, and transforms it into a vector whose components are in strictly increasing order,

$$\lambda_c \leftarrow \begin{cases} x_1 & \text{if } c = 1 \\ x_1 + \sum_{i=2}^c \exp(x_i) & \text{if } c > 1 \end{cases} \quad (8)$$

This transformation is invertible and differentiable and thus does not obstruct gradient-based inference of model parameters (App. B).

Ordered Beta. We model Goldstein values p_n as Beta random variables. For intensity class $z_n = c$, we assume $p_n \sim \text{Beta}(\omega_c^{(p)}, \kappa_c^{(p)})$, where $\omega_c^{(p)} \in (0, 1)$ is the mode and $\kappa_c^{(p)} > 0$ is the concentration parameter. We impose an ordering on the modes by positing a transformed OrderedNormal prior:

$$S^{-1}(\omega^{(p)}) \sim \text{OrderedNormal}_C(\mu, \sigma) \quad (9)$$

where $\omega^{(p)} = (\omega_1^{(p)}, \dots, \omega_C^{(p)})$ is the C -dimensional ordered vector of modes and $S^{-1}(\cdot)$ is the element-wise inverse sigmoid function. To ensure that each class-conditional Beta distribution is unimodal around its mode $\omega_c^{(p)}$, we then mandate that the concentration parameter $\kappa_c^{(p)}$ is greater than 2 by imposing a shifted Gamma prior.

$$(\kappa_c^{(p)} - 2) \stackrel{\text{iid}}{\sim} \text{Gamma}(k, \eta) \quad (10)$$

where k and η are the shape and rate parameters. We note that while the elements of $\omega^{(p)}$ are ordered, those of $\kappa^{(p)}$ are not.

Ordered Zero-inflated Geometric. We model the casualty counts q_n as Zero-inflated Geometric random variables. For intensity class $z_n = c$, we assume $q_n \sim \text{Zero-infl. Geom.}(\delta_c^{(q)}, \mathbf{b}_c^{(q)})$ where $\delta_c^{(q)} \in (0, 1)$ is the “gate” parameter—i.e., the inflated probability of sampling a zero—and $\mathbf{b}_c^{(q)} \in$ ered more “intense” than one with military casualties. In this work, however, we focus just on learning scales from observations that are naturally cardinal.

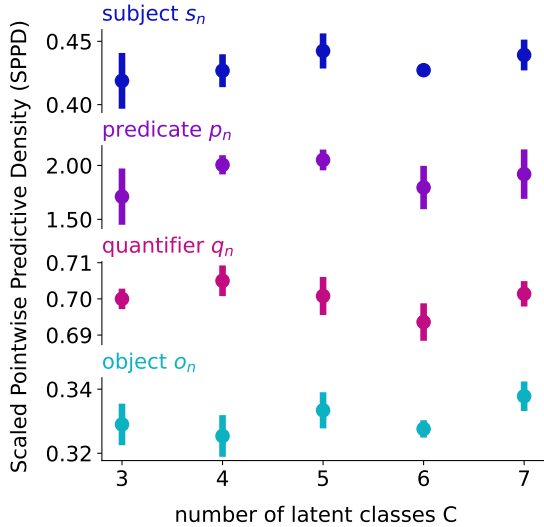


Figure 3: Determining the optimal number of latent classes C . The values show the mean and standard deviation of the scaled pointwise predictive density (SPPD) on the held-out dataset over 5 random seeds. A model with $C = 5$ latent classes fits the data best overall.

$(0, 1)$ is the success probability parameter under the standard Geometric distribution. We impose an ordering on both C -dimensional vectors of parameters $\delta^{(q)}$ and $\mathbf{b}^{(q)}$ using the same transformed Ordered Normal prior as in the previous subsection:

$$S^{-1}(\delta^{(q)}) \sim \text{OrderedNormal}_C(\mu, \sigma) \quad (11)$$

$$S^{-1}(\mathbf{b}^{(q)}) \sim \text{OrderedNormal}_C(\mu, \sigma) \quad (12)$$

Importantly though, we reverse the ordering of λ such that for all $c \in \{1, \dots, C\} : \lambda_c < \lambda_{c-1}$. The reason is that higher $\delta_c^{(q)}$ and $b_c^{(q)}$ correspond, on average, to lower values sampled from the Zero-inflated Geometric and thus lower casualty counts and lower event intensity.

4.2 Posterior Inference

We implement the model using the probabilistic programming framework Pyro (Bingham et al., 2018). Pyro provides an `OrderedTransform` function that implements eq. (8). We approximate the posterior distribution of our model’s parameters using the No-U-Turn Sampler (NUTS; Homan and Gelman, 2014), a variant of Hamiltonian Monte Carlo. NUTS is gradient-based and requires the parameters to be continuous. However, we explicitly model our latent variable z_n to be ordinal. To solve this problem, Pyro provides `parallel_enumeration` which marginalizes out discrete variables numerically during inference.

We refer to all continuous parameters of our model as $\theta = \{\boldsymbol{\pi}^{(z)}, \boldsymbol{\Pi}^{(s)}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\kappa}^{(p)}, \boldsymbol{\delta}^{(q)}, \mathbf{b}^{(q)}, \boldsymbol{\Pi}^{(o)}\}$ and to the full dataset of event tuples as X . NUTS produces samples of θ from the posterior $\theta^{(t)} \sim p(\theta | X)$. We can further sample the ordinal latent $z_n^{(t)} \sim p(z_n | \theta^{(t)}, X)$ using Pyro’s `infer_discrete`. Based on the samples $\{z_n^{(t)}\}_{t=1}^T$ per event n , we can then compute a point estimate of the event’s ordinal intensity either by taking the mean $\bar{z}_n = \frac{1}{T} \sum_{t=1}^T z_n^{(t)}$ or the mode (i.e., the most frequently sampled class) \hat{z}_n .

Label Switching. It is typically not meaningful to compute \bar{z}_n for mixture and admixture models due to the problem of label switching (Stephens, 2000), where the labels of the latent classes may switch between Markov chain Monte Carlo (MCMC) iterations. However, in the proposed model, the ordering transformation presented in eq. (8) represents an identifiability constraint that fixes the meaning of $z_n = c$ —i.e., c is the class that is more intense than $c - 1$ and less intense than $c + 1$. This alleviates the problem of label switching, and permits us to meaningfully average posterior samples $z_n^{(t)}$.

Practical Details. Our model has $13 \times C$ trainable parameters, where C is the number of latent classes. When fit to the full dataset introduced in §3, our implementation generates 1,000 posterior samples, for $C = 5$, on a CPU with 64 GB of RAM in less than 10 minutes. Throughout we set hyperparameters to uninformative values: $\mu = -1.0$, $\sigma = 1.0$, $k = 1.0$ and $\eta = 1.0$.

5 Intrinsic Evaluation: Imputation

To evaluate the fit of our model, we conduct a series of predictive experiments. We randomly split the $N = 112,089$ events into a 70% training and 30% held-out set. We fit the model on the training set X and approximate the posterior distribution of continuous parameters θ based on $T = 1,000$ NUTS samples after discarding the first 200 burn-in samples. For evaluation on the held-out set Y , we report scaled pointwise predictive density (SPPD), mean squared error and weighted F1 score as detailed in App. A.

5.1 Number of Latent Classes C

The number of classes C that our latent variable z_n ranges over can be regarded a hyperparameter. To identify the optimal setting of C , we repeatedly

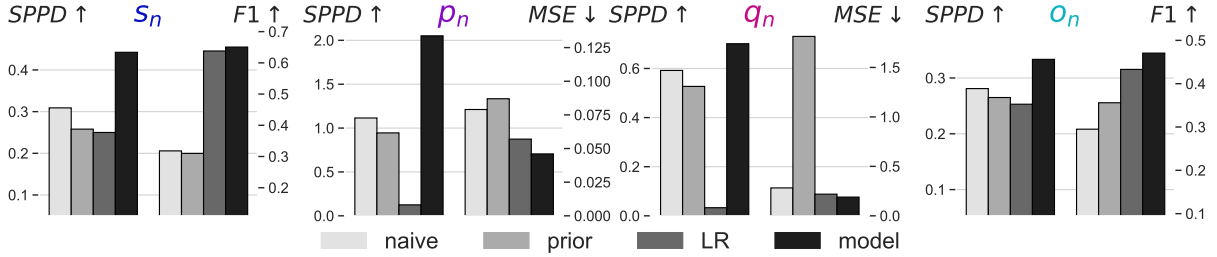


Figure 4: Intrinsic evaluation on the held-out set: we impute removed values y_n , e.g., $\{p_n\}$, of one site based on values y_n^{remain} of remaining sites, e.g., $\{s_n, q_n, o_n\}$. We present mean results of scaled pointwise predictive density (*SPPD*) and error metrics (*weighted F1 / MSE*) over 5 runs. We compare three variants of our model with $C = 5$ latent classes: a *naive* baseline fitted only to the single site that is removed at test time; an unfitted version, whose parameters are randomly drawn from the *prior*, and a fitted variant of the proposed *model*. In addition, we train four linear regression (*LR*) models to directly predict y_n from y_n^{remain} .

fit the model on the training set and evaluate it on the held-out set. We do so for each class setting in $C = [3, \dots, 7]$ and find that $C = 5$ yields the best model fit overall, as shown in Fig. 3.

5.2 Imputation Task

Procedure. Next, we conduct a data imputation task. On the held-out set, we remove one observed site from all event tuples and refer to it as y_n — e.g., we remove all Goldstein values and obtain $y_n = \{p_n\}$. We use the remaining three-way event tuples $y_n^{\text{remain}} = \{s_n, q_n, o_n\}$ jointly with the posterior of θ to make predictions of y_n . In particular, we draw samples from the posterior predictive distribution $\hat{y}_n^{(t)} \sim p(y_n | \theta^{(t)}, y_n^{\text{remain}})$ and compute their mean $\hat{y}_n = \frac{1}{T} \sum_{t=1}^T \hat{y}_n^{(t)}$ to get point estimates of y_n .

Baselines. We compare the proposed model against three baselines, where the first two baselines are simple and we expect our proposed model to outperform them. For the first baseline, we fit four versions of the proposed model to only one of the observed sites $\{s_n, p_n, q_n, o_n\}$. This naive baseline effectively just learns the empirical histogram of each site in the training set, and uses the relevant one to make predictions for y_n — we refer to this as *naive*. For the second baseline, we use an unfitted version of the proposed model, whose parameters have been sampled from the *prior*. For the last baseline, we train four linear regression models (*LR*) to predict y_n directly from y_n^{remain} .

Results. Fig. 4 reports the results of the imputation experiments, where the proposed model performs comparatively well, obtaining substantially better predictive performance over both the naive and non-naive baselines.

6 Extrinsic Evaluation: Time Series

In the last section, we intrinsically evaluated the proposed model by having it impute held-out data. In this section, we consider extrinsic evaluations of the proposed model, which seek to evaluate it on downstream predictive tasks for which it was not explicitly designed. In this section, we rely on features of events coded by NAVCO which our model does not access during training. In particular, NAVCO codes for the location ℓ_n and the timestamp τ_n of each event. We discretize time so that each event is associated with a specific month m_n . Augmented with these extra characteristics, as well as with the mean intensity inferred by our model, each event n has the following attributes: $\{s_n, p_n, q_n, o_n, \ell_n, m_n, \bar{z}_n\}$. In all of the tasks in this section, we split events according to their location $\ell_n = \ell$ and examine monthly-aggregated time series of the cardinal and ordinal quantities $\{p_n, q_n, \bar{z}_n\}$. In particular, we ask two questions: 1) whether \bar{z}_n adds predictive information for forecasting p_n and q_n , and 2) whether the model-based measure of intensity \bar{z}_n correlates with Google Trends more than just p_n and q_n .

Event Time Series. We construct time series in the following way. We first run inference on a training dataset X to obtain posterior samples of the parameters $\{\theta^{(t)}\}_{t=1}^T$. We then use the trained parameters to obtain intensity scores $\{z^{(t)}\}_{t=1}^T$ for both events in the training set as well as events in a test set that the model did not access. We then compute \bar{z}_n for each n and re-scale the \bar{z}_n estimates to cover a $[0, 1]$ range as we did with Goldstein values. For each location ℓ and month m appearing in the data, we then construct the

country ℓ	forecasting Goldstein time series \bar{p}_ℓ					forecasting casualty count time series \bar{q}_ℓ				
	$\bar{p}_\ell \rightarrow \bar{p}_\ell$	$\bar{p}_\ell, \bar{q}_\ell \rightarrow \bar{p}_\ell$	Grang	$\bar{p}_\ell, \bar{z}_\ell \rightarrow \bar{p}_\ell$	Grang	$\bar{q}_\ell \rightarrow \bar{q}_\ell$	$\bar{p}_\ell, \bar{q}_\ell \rightarrow \bar{q}_\ell$	Grang	$\bar{q}_\ell, \bar{z}_\ell \rightarrow \bar{q}_\ell$	Grang
Egypt	3.33	3.15	0.30	3.14	0.19	0.22	0.15	0.79	0.14	0.35
Iraq	4.85	4.82	0.05	4.82	0.05	2.91	2.82	0.05	2.82	0.14
Syria	2.96	3.10	0.29	3.10	0.35	0.56	0.49	0.28	0.50	0.10
Yemen	5.09	5.12	0.15	5.12	0.46	1.41	1.02	0.15	1.02	0.02

Table 2: Forecasting: we test how much predictive information our intensity time series \bar{z}_ℓ holds on future developments of Goldstein values \bar{p}_ℓ and casualty counts \bar{q}_ℓ . In most cases, the vector autoregression (VAR) outperforms the autoregression in terms of mean squared error (MSE) (results reported are 10^{-1}) on the held-out set. We find that \bar{z}_ℓ holds information on both \bar{p}_ℓ and \bar{q}_ℓ as including \bar{z}_ℓ as an additional time series performs en par with including \bar{p}_ℓ or \bar{q}_ℓ throughout the VAR experiments. If the p value of the corresponding Granger (Grang) test is small, we can reject the null hypothesis that the additional time series does *not* add predictive information.

posterior average intensity class $\bar{z}_{\ell,m}$:

$$\bar{z}_{\ell,m} = \frac{\sum_{n=1}^N \bar{z}_n \mathbb{1}(\ell_n = \ell) \mathbb{1}(m_n = m)}{\sum_{n=1}^N \mathbb{1}(\ell_n = \ell) \mathbb{1}(m_n = m)} \quad (13)$$

For a given location ℓ , we refer to the full time series over months as \bar{z}_ℓ and linearly interpolate the entries m corresponding to months for which there is no data. We similarly aggregate the observed sites per event to obtain time series of predicates \bar{p}_ℓ and quantifiers \bar{q}_ℓ .

Autoregressive Forecasting. Does knowledge of the inferred intensity time series \bar{z}_ℓ improve forecasting of the Goldstein \bar{p}_ℓ and casualty count \bar{q}_ℓ time series? To test this, we first consider autoregressive (AR) models that forecast monthly values of each time series (e.g., $\bar{p}_{\ell,m}$) based on the previous values of only that same time series (e.g., $\bar{p}_{\ell,1:(m-1)}$). We then consider vector autoregressive (VAR) models that use multiple time series (e.g., $\bar{p}_{\ell,1:(m-1)}$ and $\bar{q}_{\ell,1:(m-1)}$) to form the same predictions. When incorporating values from our latent intensity time series \bar{z}_ℓ , we must be careful to avoid test set leakage. To do so, we fit our model to a subset of data $X_{\setminus \ell}$ that excludes any events in a particular location ℓ . We obtain posterior samples $\{\theta_{\setminus \ell}^{(t)}\}_{t=1}^T$ by fitting the proposed model to $X_{\setminus \ell}$. We then use these parameters to obtain \bar{z}_n for all events n in both the training $X_{\setminus \ell}$ and test set X_ℓ .

Before fitting any (vector) autoregressive models and performing the forecasting experiments, we apply a first-order differencing transform to all time series to remove potential linear trends, and verify each time series' stationarity using an Augmented Dickey-Fuller test (Fuller, 1976). We then fit an autoregressive (AR) model to each individual time-

series and a vector autoregressive (VAR) model to all pairs of time series. We determine their optimal orders (lag) in months using Bayesian Information Criterion (BIC), and use cross validation to measure held-out forecasting error across 24 folds. In Tab. 2, we report results in mean squared error (MSE) on the held-out sets averaged over all 24 folds. Indeed, our results suggest that \bar{z}_ℓ contains (predictive) information on both \bar{p}_ℓ and \bar{q}_ℓ . For instance, in the VAR experiments for forecasting \bar{p}_ℓ , we may replace the additional time series \bar{q}_ℓ with \bar{z}_ℓ without any drop in performance. We also report Granger tests (Granger, 1969) which test the null hypothesis that forecasting a variable (e.g., \bar{p}_ℓ) using only its own history is no less accurate than also using an additional variable's history (e.g., \bar{z}_ℓ).

Descriptive Analysis. A growing body of work analyzes the correlation between shifts in online behavior or media attention and conflict intensity (Chykina and Crabtree, 2018; Timoneda and Wibbels, 2022). Following this line of work, we download time series of Google search keywords using the Google Trends Anchorbank (G-TAB) (West, 2020). We constrain searches to the category of "World News" and use the country name of location ℓ as the search keyword. In contrast to the forecasting setting, we fit our model to the entire dataset X , including events associated with the respective location, and obtain an intensity time series \bar{z}_ℓ . Fig. 5 shows a comparison between four time series: the Goldstein \bar{p}_ℓ , casualty counts \bar{q}_ℓ , our latent \bar{z}_ℓ and the Google trends time series for Syria between 2004 and 2013.³ We observe that

³Fisher (2012) reports a correlation between Google search volume and activities in the Syrian Civil War for this period.

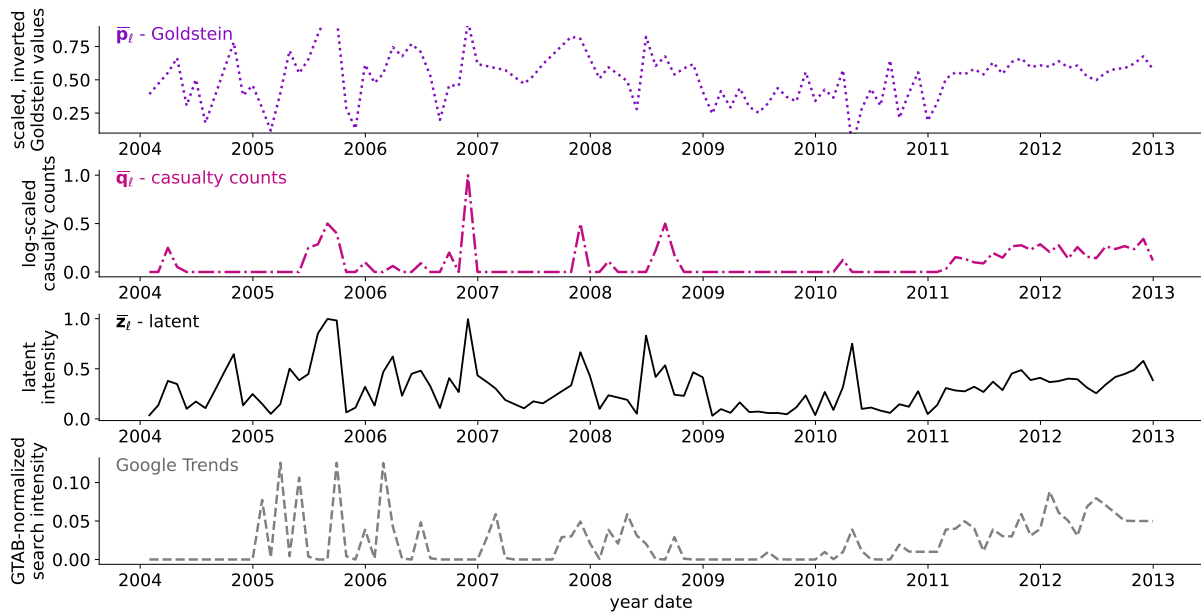


Figure 5: Time series of events in Syria between 2004 and 2013, all values averaged by month. We observe that our intensity time series \bar{z}_ℓ smoothly interpolates between the Goldstein \bar{p}_ℓ and casualty counts time series \bar{q}_ℓ . We look at Google trends as an external correlate for (perceived) conflict intensity and report correlations in Fig. 6.

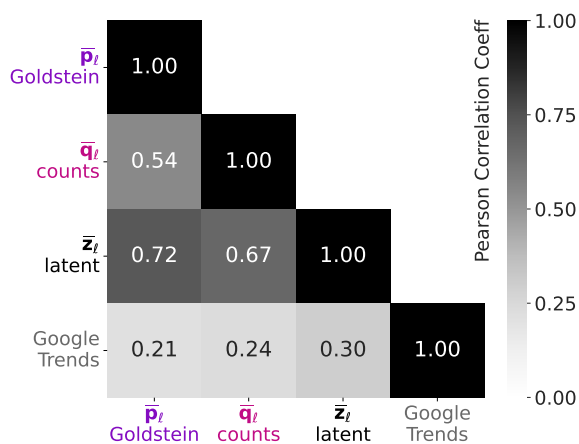


Figure 6: Pearson correlation between time series displayed in Fig. 5. Our latent intensity time series \bar{z}_ℓ is positively correlated with the Goldstein \bar{p}_ℓ and casualty counts time series \bar{q}_ℓ . Google trends are more strongly correlated with \bar{z}_ℓ than with \bar{p}_ℓ and \bar{q}_ℓ .

\bar{z}_ℓ nicely interpolates between the time series of \bar{p}_ℓ and \bar{q}_ℓ . It captures larger trends and fluctuations in both time series, while not exactly mirroring either. Further, in Fig. 6, we report Pearson correlation coefficients between all pairs of time series. While \bar{p}_ℓ and \bar{q}_ℓ are positively correlated, \bar{p}_ℓ is more strongly correlated with \bar{z}_ℓ than with \bar{q}_ℓ . Moreover, Google trends are more strongly correlated with \bar{z}_ℓ than either \bar{p}_ℓ or \bar{q}_ℓ . We hypothesize this may be due

2004 is the earliest date to query Google trends.

to the additional information on the subject s_n and object types o_n that z_n encodes, and which may contribute to how much attention is paid to different conflicts, as measured by Google Trends.

7 Related Work

There are a number of relevant papers that broadly seek to measure “latent concepts” (Douglass et al., 2022) pertaining to international relations, particularly using event data. Terechshenko (2020) uses a Bayesian item-response theory model to learn ordinal conflict intensity levels from observed event types. O’Connor et al. (2013) present an unsupervised, probabilistic topic model to learn an ontology of event data and use the Goldstein scale to evaluate it. Schein et al. (2015, 2016, 2019) decompose four-way tensors (senders, receivers, actions, time steps) to infer latent classes of CAMEO-coded events. Stoehr et al. (2023b) build on those models by further imposing an ordering on their latent space, which captures conflict-cooperation intensity. Another line of work models friend-enemy relationship trajectories using neural network-based (Han et al., 2019) or hidden Markov model-based (Chaturvedi et al., 2017) approaches. There also exists work on signed network representations of relationships that are extracted from text (Srivastava et al., 2016; Choi et al., 2016) or Wikipedia conflict articles (Stoehr et al., 2021).

8 Future Work

Assuming civilian casualties are “more intense” than military casualties, we could impose an additional ordering on object types o_n . We could also condition observed sites on each other—e.g., casualty counts q_n are conditionally dependent on o_n or even p_n under the model. We plan to further incorporate multiple latent variables to model multi-dimensional intensity concepts. Future models could also condition on location and include a temporal component to account for how surprisal may affect the perceived intensity of an event. We also note that the latent variable model presented in this work could also be extended to a more general framework for learning interpretable, ordinal scales from a set of mixed-type data that include cardinal or ordinal observations. We plan to explore generalizations of the proposed model and applications beyond of international relations.

Acknowledgments

We would like to thank Giuseppe Russo and Benjamin Radford for helpful discussions. Moreover, we acknowledge the feedback received from the anonymous reviewers, the NLP groups at ETH Zurich and the DLAB at EPFL. NS acknowledges support from the Swiss Data Science Center (SDSC) fellowship. LTH acknowledges support from the Michael Athans fellowship fund.

Limitations

We discuss limitations of our modeling assumptions in §8. They are based on prior work in political science such as the CAMEO ontology and the Goldstein scale. On this account, they may replicate or potentially introduce biases. Hyperparameter search, settings and implementation details are provided in §4.2 and §5. All NAVCO event descriptions are limited to English language, but do not disclose individuals.

Impact Statement

We emphasize that our models are intended for research, analysis and monitoring purposes. They should not be blindly deployed for automatized decision-making processes. The notion of conflict intensity is intrinsically hard to quantify: it is strongly dependent on socio-cultural background and subjective experience.

References

- John Beiler. 2016. [Creating a real-time, reproducible event dataset](#). *arXiv*, 1612.00866.
- Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. 2018. [Pyro: Deep universal probabilistic programming](#). *Journal of Machine Learning Research*.
- Doug Bond, Joe Bond, Churl Oh, 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.
- Elizabeth Boschee, Jennifer Lautenschlager, Sean O’Brien, Steve Shellman, James Starz, and Michael Ward. 2015. [ICEWS coded event data](#).
- Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. 2017. [Unsupervised learning of evolving relationships between literary characters](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).
- Stephen Chaudoin, Zachary Peskowitz, and Christopher Stanton. 2017. [Beyond zeroes and ones: The intensity and dynamics of civil conflict](#). *The Journal of Conflict Resolution*, 61(1):56–83.
- Erica Chenoweth, Jonathan Pinckney, and Orion Lewis. 2018. [Days of rage: Introducing the NAVCO 3.0 dataset](#). *Journal of Peace Research*, 55(4):524–534.
- Eunsol Choi, Hannah Rashkin, Luke Zettlemoyer, and Yejin Choi. 2016. [Document-level sentiment inference with social, faction, and discourse context](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 333–343.
- Volha Chykina and Charles Crabtree. 2018. [Using Google trends to measure issue salience for hard-to-survey populations](#). *Socius: Sociological Research for a Dynamic World*, 4:237802311876041.
- Rex Douglass, Thomas Leo Scherer, Andrés Gannon, Erik Gartzke, Jon Lindsay, Shannon Carcelli, Jonathan Wilkenfeld, David M. Quinn, Catherine Aiken, Jose Miguel Cabezas Navarro, Neil Lund, Egle Murauskaite, and Diana Partridge. 2022. [Introducing the ICBe dataset: Very high recall and precision event extraction from narratives about international crises](#). *arXiv*, 2202.07081.
- Max Fisher. 2012. [Google trends: The moment Syria’s ‘revolution’ became a ‘civil war’](#). *Washington Post*.
- Wayne A. Fuller. 1976. *Introduction to Statistical Time Series*. Wiley.
- Andrew Gelman, Jessica Hwang, and Aki Vehtari. 2014. [Understanding predictive information criteria for Bayesian models](#). *Statistics and Computing*, 24(6):997–1016.

- Andrew Gelman, Xiao-Li Meng, and Hal Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760.
- Joshua Goldstein. 1992. A conflict-cooperation scale for WEIS events data. *The Journal of Conflict Resolution*, 36(2):369–385.
- Clive Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.
- Xiaochuang Han, Eunsol Choi, and Chenhao Tan. 2019. No permanent friends or enemies: Tracking relationships between nations from news. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1, pages 1660–1676.
- Matthew D. Homan and Andrew Gelman. 2014. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.
- Gary King 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.
- Kalev Leetaru and Philip Schrodt. 2013. GDELT: Global data on events, location, and tone. In *ISA Annual Convention*.
- Orion A. Lewis, Erica Chenoweth, and Jonathan Pinckney. 2016. Nonviolent and violent campaigns and outcomes 3.0: Effects of tactical choices on strategic outcomes codebook.
- Charles McClelland. 1984. World event/interaction survey (WEIS) project, 1966-1978: Archival version.
- Brendan O’Connor, Brandon Stewart, and Noah Smith. 2013. Learning to extract international relations from political context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1094–1104.
- Aaron Schein, Scott W. Linderman, Mingyuan Zhou, David M. Blei, and Hanna Wallach. 2019. Poisson-randomized gamma dynamical systems. In *Advances in Neural Information Processing Systems*.
- Aaron Schein, John Paisley, David M. Blei, and Hanna Wallach. 2015. Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1045–1054.
- Aaron Schein, Mingyuan Zhou, David M. Blei, and Hanna Wallach. 2016. Bayesian Poisson Tucker decomposition for learning the structure of international relations. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 2810–2819.
- Philip Schrodt. 2008. Kansas event data system (KEDS). In *AAAS Center for Scientific Responsibility and Justice*.
- Philip Schrodt. 2012. CAMEO: Conflict and mediation event observations event and actor codebook. *Parus Analytics*.
- Philip Schrodt. 2019. Stuff I tell people about event data.
- Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2807–2813.
- Matthew Stephens. 2000. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.
- Niklas Stoehr, Ryan Cotterell, and Aaron Schein. 2023a. Sentiment as an ordinal latent variable. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.
- Niklas Stoehr, Lucas Torroba Hennigen, Samin Ahabab, Robert West, and Ryan Cotterell. 2021. Classifying dyads for militarized conflict analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Niklas Stoehr, Benjamin J. Radford, Ryan Cotterell, and Aaron Schein. 2023b. The Ordered Matrix Dirichlet for state-space models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*.
- Zhanna Terechshenko. 2020. Hot under the collar: A latent measure of interstate hostility. *Journal of Peace Research*, 57(6):764–776.
- Joan C. Timoneda and Erik Wibbels. 2022. Spikes and variance: Using Google trends to detect and forecast protests. *Political Analysis*, 30(1):1–18.
- Robert West. 2020. Calibration of Google trends time series. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2257–2260.
- Mian Zhong, Shehzaad Dhuliawala, and Niklas Stoehr. 2023. Extracting victim counts from text. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

A Evaluation Metrics

To evaluate the posterior predictive distribution $p(y_n | \theta^{(t)})$, we consider a scaled variant of the log pointwise predictive density (LPPD; Gelman et al., 1996, 2014), which we term *SPPD*:

$$\exp\left(\frac{1}{N} \sum_{n=1}^N \log\left(\frac{1}{T} \sum_{t=1}^T p(y_n | \theta^{(t)})\right)\right) \quad (14)$$

The term inside the log, $\frac{1}{T} \sum_{t=1}^T p(y_n | \theta^{(t)})$, computes a Monte Carlo approximation to the posterior predictive density for a given y_n . The rest, $\exp\left(\frac{1}{N} \sum_{n=1}^N \log(\cdot)\right)$, then computes the geometric mean of the pointwise densities.

We also compute point estimates via the posterior predictive mean $\frac{1}{T} \sum_{t=1}^T \mathbb{E}[y_n | \theta^{(t)}]$ which allows comparing predicted and true values based on error metrics like weighted F1 score or mean-squared error (MSE).

B Inverse of the Ordered Normal

We define the Ordered Normal distribution with the help of an ordering transformation in eq. (8). This transformation is a smooth bijection since its inverse is given by:

$$x_c \leftarrow \begin{cases} \lambda_1 & \text{if } c = 1 \\ \log(\lambda_c - \lambda_{c-1}) & \text{if } c > 1 \end{cases} \quad (15)$$

Note that for all c , $\lambda_c > \lambda_{c-1}$, so the log is well-defined.

CAMEO actor code	meaning	group 4 mapping
REB	rebels	military
MIL	military	military
GOV	government	government
ETH	ethnic	civilian
REL	religious	civilian
COP	police forces	military
JUD	judiciary	political
OPP	political opposition	political
LLY	regime loyalists	government
ACT	activists	political
NON	non-aligned third party	military
SPY	state intelligence	military
UAF	unidentified armed forces	military
UNS	unidentified unarmed non-state actors	civilian
NGO	non-governmental organisation	political
BUS	business	civilian
CVL	civilian group	civilian
IND	civilian individual	civilian
EDU	educators	civilian
STU	students	civilian
YTH	youth	civilian
ELI	elites	civilian
LAB	labour	civilian
LEG	legislature	political
PTY	political party	political
MED	media	civilian
REF	refugees	civilian
IGO	inter-governmental	political
NGM	non-governmental movement	political
MNC	multinational cooperation	civilian
INT	international actors	political
TOP	top officials	political
MID	mid-lower level officials	political
HAR	hardliners	political
MOD	moderates	political

Table 3: CAMEO actor types to group 4 mapping. The NAVCO columns “actor3”, “actor6” and “target3”, “target6” correspond to actor types defined by the [CAMEO actor codebook](#). We consider 33 generic CAMEO actor types and map them into one of the 4 classes “civilian”, “military”, “governmental” and “political”.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
section 9 and limitations section after conclusion
- A2. Did you discuss any potential risks of your work?
section 9, limitations section and impact statement after conclusion
- A3. Do the abstract and introduction summarize the paper’s main claims?
abstract and section 1
- A4. Have you used AI writing assistants when working on this paper?
–

B Did you use or create scientific artifacts?

section 3 and section 7

- B1. Did you cite the creators of artifacts you used?
section 3 and section 7
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
section 3, section 7 and appendix
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
section 1, section 3 and section 7
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
section 3, section 7, limitations section and impact statement after conclusion
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
section 3 and appendix
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 3, section 6 and appendix

C Did you run computational experiments?

section 4, section 5 and section 6

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
section 4, section 5 and appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

section 4, section 5 and appendix

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

section 2, section 4, section 5, section 6 and appendix

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Not applicable. did not use existing packages

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

Not applicable. Left blank.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

Not applicable. Left blank.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

Not applicable. Left blank.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

Not applicable. Left blank.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

Not applicable. Left blank.