Corpus-Level Evaluation for Event QA: The IndiaPoliceEvents Corpus Covering the 2002 Gujarat Violence

Andrew Halterman[*]
Massachusetts Institute of Technology
ahalt@mit.edu

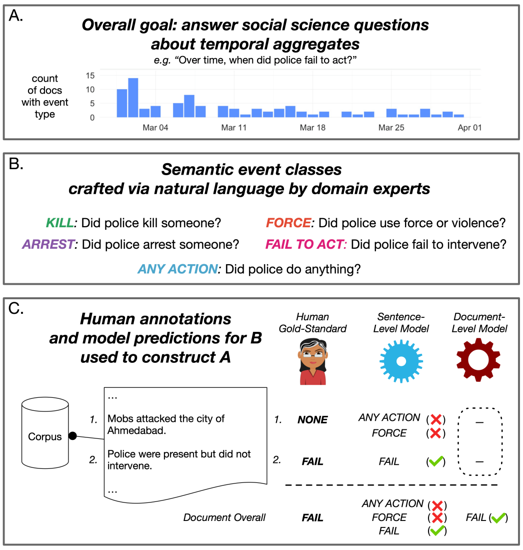
Katherine A. Keith[*]
University of Massachusetts Amherst
kkeith@cs.umass.edu

Sheikh Muhammad Sarwar[*]
University of Massachusetts Amherst
smsarwar@cs.umass.edu

Brendan O'Connor
University of Massachusetts Amherst
brenocon@cs.umass.edu

Abstract

Automated event extraction in social science applications often requires corpus-level evaluations: for example, aggregating text predictions across metadata and unbiased estimates of recall. We combine corpus-level evaluation requirements with a real-world, social science setting and introduce the INDIAPOLICEEVENTS corpus—all 21,391 sentences from 1,257 English-language *Times of India* articles about events in the state of Gujarat during March 2002. Our trained annotators read and label every document for mentions of police activity events, allowing for unbiased recall evaluations. In contrast to other datasets with structured event representations, we gather annotations by posing natural questions, and evaluate off-the-shelf models for three different tasks: sentence classification, document ranking, and temporal aggregation of target events. We present baseline results from zero-shot BERT-based models fine-tuned on natural language inference and passage retrieval tasks. Our novel corpus-level evaluations and annotation approach can guide creation of similar social-science-oriented resources in the future.



A.
Overall goal: answer social science questions
about temporal aggregates
e.g. "Over time, when did police fail to act?"
count of docs with event type
15
10
5
0
Mar 04
Mar 11
Mar 18
Mar 25
Apr 01
B.
Semantic event classes
crafted via natural language by domain experts
KILL: Did police kill someone?
FORCE: Did police use force or violence?
ARREST: Did police arrest someone?
FAIL TO ACT: Did police fail to intervene?
ANY ACTION: Did police do anything?
C.
Human annotations
and model predictions for B
used to construct A
Human Gold-Standard
Sentence-Level Model
Document-Level Model
Corpus
...
1. Mobs attacked the city of Ahmedabad.
2. Police were present but did not intervene.
...
1. NONE
ANY ACTION (✗)
FORCE (✗)
–
2. FAIL
FAIL (✓)
–
Document Overall
FAIL
ANY ACTION (✗)
FORCE (✗)
FAIL (✓)
FAIL (✓)

Figure 1: Motivation (A-B) and procedures (B-C) for this paper: **A.** Social scientists often use text data to answer substantive questions about temporal aggregates. **B.** To answer these questions, domain experts use natural language to define semantic event classes of interest. **C.** Our INDIAPOLICEEVENTS dataset: Humans annotate *every* sentence in the corpus in order to evaluate whether a system achieves full recall of relevant events. In production, computational models run B's queries to classify or rank sentences or documents, which are aggregated to answer A.

1 Introduction

Understanding the actions taken by political actors is at the heart of political science research: How do actors respond to contested elections (Daxecker et al., 2019)? How many people attend protests (Chenoweth and Lewis, 2013)? Which religious groups are engaged in violence (Brathwaite and Park, 2018)? Why do some governments try to prevent anti-minority riots while others do not (Wilkinson, 2006)? In the absence of official records, social scientists often turn to news data to extract the actions of actors and surrounding events. These news-based event datasets are often constructed by hand, requiring large investments of time and money and limiting the number of researchers who can undertake data collection efforts.

Automated extraction of political events and actors is already prominent in social science (Schrodt et al., 1994; King and Lowe, 2003; Hanna, 2014; Hammond and Weidmann, 2014; Boschee et al., 2015; Beieler et al., 2016; Osorio and Reyes, 2017) and is increasingly promising given recent gains in information extraction (IE), the automatic conversion of unstructured text to structured datasets (Grishman, 1997; McCallum, 2005; Grishman, 2019). While social scientists and IE researchers have over-

[*] Indicates joint first-authorship.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4240–4253
August 1–6, 2021. ©2021 Association for Computational Linguistics

lapping interests in evaluating event extraction systems, social scientists have particular needs that have so far been under-addressed by the computer science IE research community.

Figure 1A shows a common goal of social scientists: answering aggregate substantive questions from corpora such as, *"Over time, when did police fail to act?"* which could be measured by, for example, the daily count of newspapers mentioning the event class over time. For these types of questions, social scientists predominantly want very high recall methods because often the events of interest are sparse or their substantive conclusions depend on identifying *every* event in a corpus.[1]

In contrast to this corpus-level focus, much of current IE research has focused on distinct subtasks such as entity linking, relation extraction, and coreference resolution.[2] Furthermore, all widely used event datasets (e.g. ACE, FrameNet, ERE, or KBP; Aguilar et al. 2014) are typically curated at the *ontology* level—attempting to cover a selected set of event types—but have little consideration of the *corpus* level—annotated documents are not necessarily a substantively meaningful sample of the broader corpora from which they are drawn. We try to address these evaluation shortcomings in this paper.

In addition to corpus-level recall, social scientists are often interested in using off-the-shelf models that are easily extensible to their domain questions. Fortunately, recent NLP research has seen a paradigm shift from structured semantic and event representations (Abend and Rappoport, 2017; Aguilar et al., 2014) which are limited by their predefined schemas, to directly using natural language to encode semantic arguments (*QA-SRL*; He et al. 2015; Stanovsky et al. 2016; FitzGerald et al. 2018; Roit et al. 2020) and events (Levy et al., 2017; Liu et al., 2020; Du and Cardie, 2020). In this paper, we also use natural language questions to annotate and model the event classes in our dataset, not only facilitating ease of annotation, but also allowing for the evaluation of zero-shot natural language inference and information retrieval models for the tasks.

To address these social science desiderata, we present the INDIAPOLICEEVENTS corpus[3] which has the following useful properties:

- **Social science relevance.** Our dataset consists of all 21,391 sentences from all 1,257 *Times of India* articles about events in the state of Gujarat during March, 2002—a period that is of deep interest to political scientists due to widespread Hindu–Muslim violence (Dhattiwala and Biggs, 2012; Berenschot, 2012; Basu, 2015). We focus on the actions of a single entity type, *police*, because of extensive substantive research on police actions during the Gujarat violence (Varadarajan, 2002). Our choice of location, actors, and event types are motivated by Wilkinson (2006)—political science work which created a hand-coded event dataset from newspapers about communal violence events in India from 1950-1995.
- **Corpus-level full-recall.** Unlike most previous event evaluation datasets, our annotators read *every* document in our corpus (that match a loose spatiotemporal filter; §4.1). This requires substantially more annotation work compared to a more targeted filter to select documents to annotate (e.g. matching via keywords), but eliminates a potential source of evaluation bias compared to alternative document retrieval data collection approaches (Grossman et al., 2016), and allows for full-recall evaluation of end-to-end event extraction systems.
- **Document-level context.** Our annotators read the context of an entire document to provide answers for each question on each sentence. We then aggregate these sentence-level answers to make document-level inferences. This allows us to accurately label sentences with anaphora or context-specific meaning.
- **Natural language event specification and zero-shot model evaluation.** In constructing our dataset, we gather annotations via a natural question-answer format because it allows for easily specifying constraints on arguments (e.g. *police* being the agent). Additionally, it allows for specifying event predicates not covered within the ontologies of current structured semantic representations, or with additional hard-to-specify

[1] In some studies, researchers rely on an assumption that events are missing at random, but others depend on knowing whether an event occurred at least once.

[2] The first five Message Understanding Conferences (MUC) required participants to submit complete systems to fill event templates; however, starting with MUC-6 and subsequent ACE and KBP tasks, information extraction was broken into distinct modules (Grishman and Sundheim, 1996; Grishman, 2019).

[3] Dataset, source code, and appendix are provided at http://slanglab.cs.umass.edu/IndiaPoliceEvents/ and https://github.com/slanglab/IndiaPoliceEvents.

semantic phenomena—e.g. "Did police fail to act?" or when political actors do *not* take an action, which is very important to political scientists (e.g. Wilkinson (2006)). This format also allows us to evaluate zero-shot natural language inference and information retrieval models.

- **High-quality annotators who provide uncertainty explanations.** We hire and train political science undergraduate students as annotators to ensure quality control, retraining annotators over a period of several months with training videos, two hour-long live meetings, and individual annotator feedback before producing our final dataset. Our annotators also provide free-text explanations for instances in which they are uncertain about the answer. These rationales are important given the recent attention to propagating annotator uncertainty in downstream NLP tasks (Dumitrache et al., 2018; Paun et al., 2018; Pavlick and Kwiatkowski, 2019; Keith et al., 2020) and social scientists' interest in quantifying uncertainty (King, 1989; Wallach, 2018).

In the remainder of this paper, we use our dataset for three levels of evaluation: sentence-level classification, ranking of documents to reduce manual reading time, and constructing temporal aggregates useful to social scientists (§3). We describe in detail our annotation and dataset creation process (§4), provide baseline models (§5), and evaluate their performance on all three tasks (§6).

2 Related Work

NLP and IR for police activity. Natural Language Processing (NLP) and Information Retrieval (IR) have been used for analysis of other police activity such as identifying victims of police fatalities from news articles (Keith et al., 2017; Nguyen and Nguyen, 2018; Sarwar and Allan, 2019); extracting eye-witness event types from Twitter including police activity and shootings (Doggett and Cantarero, 2016); detecting dialogue acts from police stops (Prabhakaran et al., 2018); and computational analysis of degree of respect in police officers' language (Voigt et al., 2017).

Political event extraction. Automated event extraction in social science is generally performed using dictionary methods and a set of substantively motivated event types and actor categories (Schrodt et al., 1994; Gerner et al., 2002; Beieler et al., 2016; Boschee, 2016; Radford, 2016; Brathwaite and Park, 2018; Liang et al., 2018). Other work uses supervised learning to infer events such as conflict or cooperation (Beieler, 2016) and protests (Hanna, 2017). While some have attempted to induce event types without supervision (O'Connor et al., 2013; Huang et al., 2016), most social science applications of event extraction require substantial human input either through constructing keyword lists, or annotating texts to train classifiers.

Recall-focused IR. TREC's total-recall track (Grossman et al., 2016) is inspired by real-world recall-focused applications from law, medicine, and oversight (McDonald et al., 2018). However, the track's datasets are not typically focused on events and assume documents are collected through interacting with a system. Other work has focused on methods for truncating ranked lists that minimize the risk of viewing non-relevant documents (Arampatzis et al., 2009; Lien et al., 2019), but this line of work does not evaluate on semantic retrieval of event classes.

3 Three Levels of Tasks

In order to answer substantive social science questions, for example *"Does variation in party control of state government affect whether police failed to intervene in ethnic conflict?"* (Wilkinson, 2006), social scientists often need to gather counts of events (e.g. "police failed to intervene") from text when official government records are lacking. Ideally, a social scientist could use automatic information extraction methods (Cowie and Lehnert, 1996; McCallum, 2005; Grishman, 2019) to transform unstructured text into a structured database that would be useful in a quantitative analysis. Yet, even state-of-the-art information extraction systems often give less than perfect accuracy, so social scientists must still manually analyze large portions of their corpus in order to extract events of interest. This quantitative research process motivates the following three tasks which our dataset can be used to evaluate:

Task 1: Sentence classification. Although social science corpora typically consist of *documents*, it would be useful for a system to classify *sentences* that contain events of interest.[4] Highlighting relevant sentences could, for semi-automated systems, reduce a social scientist's reading time, and, for fully-automated systems, provide sentence-level evidence of the automated method's *validity*, a cru-

[4] This is closely related to extracting "explanation representations" (Thayaparan et al., 2020), "supporting facts" (Yang et al., 2018b) or "evidence sentences" (Wang et al., 2019) in the machine reading comprehension literature.

cial aspect of research in text-as-data (Grimmer and Stewart, 2013) and the broader social sciences (Drost et al., 2011). IndiaPoliceEvents allows for evaluation of sentence-level precision, recall, and F1 (§6.1).

Task 2: Document ranking. For semi-automated systems, social scientists must navigate the tradeoff between recall and manual reading time. Social scientists may rely on IR methods which present ranked lists of relevant documents (Baeza-Yates et al., 1999; Schütze et al., 2008). However, our informal interviews with social scientists suggest they want to know at what point they have read enough documents to achieve very high (95–100%) recall. In creating IndiaPoliceEvents, annotators read every single sentence in a corpus which allows for full evaluations of average precision and our newly proposed metric: the proportion of the corpus that would have to be read to achieve Recall=X (PropRead@RecallX) (§6.2).[5]

Task 3: Substantive temporal aggregates. For social scientists, NLP and IR methods are used in service of answering substantive questions from text. In addressing our running example *"Did differences in party control of state government affect whether police failed to intervene in ethnic conflict?"* a social scientist could measure how many news articles[6] discuss "police failing to intervene" each day for a given temporal span. In this setting, it would be helpful to know if changes in model performance at the sentence or document level resulted in significant differences at this aggregate level. We design IndiaPoliceEvents with the capability of evaluating these meaningful corpus-level temporal aggregates, such as the mean absolute error and Spearman rank correlation coefficient between per-day event counts of computational models and ground truth annotations (§6.3).

4 Annotations and Dataset

4.1 Corpus selection

We curate our corpus with a substantively motivated specification: it is restricted to a single authoritative news source, over a defined span of time, with articles that mention one of two locations involved in or related to the 2002 Gujarat violence.

Event Class	Pos. Sents.		Pos. Docs.	
KILL	96	(0.45%)	50	(3.98%)
ARREST	299	(1.40%)	128	(10.17%)
FAIL TO ACT	207	(0.97%)	114	(9.05%)
FORCE	222	(1.04%)	90	(7.15%)
ANY ACTION	2,073	(9.69%)	457	(36.24%)

Table 1: IndiaPoliceEvents number and percentage of positive sentences (sents.) and documents (docs.) after the adjudication round. In total, the dataset contains 21,391 sentences and 1,257 documents.

From the website of *Times of India*, an English language newspaper of record in India, we first download all news articles published in March 2002.[7] During this period, widespread communal violence occurred in India, following the death of 59 Hindu pilgrims in a train fire in the state of Gujarat. In the subsequent months, reprisal attacks were directed at mostly Muslim victims across the state (Human Rights Watch, 2002; Subramanian, 2007). In creating our annotations, we specifically focus on the actions of police during these events, since a large body of evidence points to the importance of police intervention and non-intervention in quelling or permitting ethnic violence (Human Rights Watch, 2002; Wilkinson, 2006; Subramanian, 2007). We focus on the first month of the violence in order to fit within our annotation budget. This month saw the greatest levels of violence, though violence continued for a period of months afterward.

Our final corpus consists of the subset of scraped documents published in March 2002 that include either the name of the state (*Gujarat*) or a city related to the beginning of violence (*Ayodhya*).[8] Selecting on geographical and temporal metadata is a high recall way to filter the corpus without biasing the dataset by filtering to topic or event-related keywords, thus giving a better view of the true recall of an event extraction method.

[5] We do not address the problem of estimating recall when gold-standard labels are only known for the subset of documents read so far, but IndiaPoliceEvents could be used to evaluate that task in future work.

[6] Count of news articles are often used in social science as a proxy for the true measure of the event, e.g. Nielsen (2013); Chadefaux (2014).

[7] §9 discusses copyright issues.

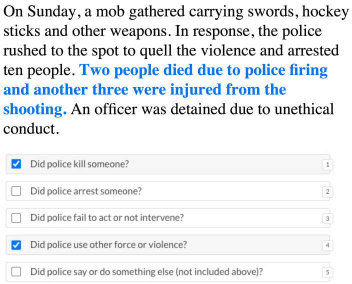
[8] Selecting documents using location-based keywords is a standard first step in political science text analysis (Mueller and Rauh, 2017). This filters to 18% of the total articles in March 2002. The precipitating event for the March 2002 violence was the burning of a train of pilgrims returning from Ayodhya.

4.2 Annotations via natural language

To collect annotations, we give annotators an entire document for context, and then ask them *natural language questions* about semantic event classes anchored on the actions of police for each sentence in that document:

- KILL: "Did police kill someone?" Lethal police violence is an important subject for social scientists (Subramanian, 2007). Example sentence: *"In Vadodara, one person was killed in police firing on a mob in the Fatehganj area."*
- ARREST: "Did police arrest someone?" Knowing when and where police made arrests and who was arrested is an important part of understanding police response to communal violence. Example sentence: *"Police officials said nearly 2,537 people have so far been rounded up in the state."*
- FAIL TO ACT: "Did police fail to intervene?" In the 2002 Gujarat violence, police were often accused of failing to prevent violence or allowing it to happen. Knowing when police were present but did not act is important for understanding the extent of this phenomenon and its potential causes (Wilkinson, 2006). Example sentence: *"The news items [...] suggest inaction by the police force [...] to deal with this situation."*
- FORCE: "Did police use force or violence?" Political scientists are interested not only when police kill but the level of force they use. Example sentence: *"Trouble broke out in Halad [...] where the police had to open fire at a violent mob."*
- ANY ACTION: "Did police do anything?" We collect annotations on all police activities, so that social scientists could, in the future, label more fine-grained event classes. Example sentence: *"In the heart of the city's Golwad area, the army is maintaining a vigil over mounting tension following [...]"*

Figure 2 shows the interface annotators see.[9] See Appendix §A for exact annotation instructions and per-question agreement rates.



On Sunday, a mob gathered carrying swords, hockey sticks and other weapons. In response, the police rushed to the spot to quell the violence and arrested ten people. **Two people died due to police firing and another three were injured from the shooting.** An officer was detained due to unethical conduct.

☑ Did police kill someone? 1

☐ Did police arrest someone? 2

☐ Did police fail to act or not intervene? 3

☑ Did police use other force or violence? 4

☐ Did police say or do something else (not included above)? 5

Figure 2: We present annotators with a highlighted sentence (blue) and its document context. Their task is to click a check-mark for the event-focused questions for which there is a positive answer in the highlighted sentence.

Following the guidelines of Pustejovsky and Stubbs (2012), we first assign each document to two annotators and then follow with an *adjudication round* in which items with disagreement are given to an additional annotator to resolve and create the gold standard. For annotators, we select undergraduate students majoring in political science (as opposed to crowdworkers) in order to approximate the domain expertise of social scientists.[10] We initially recruited and selected 12 students. After a pilot study and two rounds of training, in which we provided individual feedback to annotators via email, we selected 8 final annotators based on their performance. Each student annotated around 330 documents ($\sim$5,500 sentences) using the interface described in the Appendix, Figure A2.

Table 1 shows the prevalence of the event classes after the adjudication round. Note that some of the classes are relatively rare: of all documents, only roughly 4% have KILL and 7% have FORCE. Our annotators had fairly high inner-annotator agreement for KILL and ARREST, with Krippendorff's alpha values of 0.75 and 0.71 respectively. Other questions, such as FAIL TO ACT and "Did police use other force?" had lower agreement ($\alpha < 0.4$), indicating more difficulty and ambiguity. Full agreement rates are show in Appendix, Table A1.

4.3 Annotation uncertainty explanations

We also collect free-text *annotation uncertainty explanations* in order to analyze instances that an-

[9] While the first three classes each correspond to a single annotation question, we create FORCE and ANY ACTION by taking the union of several different questions posed to annotators, which made it easier for annotators to distinguish between different subtypes. FORCE is the union of "Did police kill someone?" and "Did police use other force or violence?". ANY ACTION is the union of four questions: "Did police kill someone?", "Did police arrest someone?", "Did police use other force or violence?", and "Did police do or say something else (not included above)?".

[10] Our annotation protocol (no. 2238) was reviewed as exempt by the University of Massachusetts Amherst's IRB office. Annotators were paid $25 per training session and a lump sum for document annotations; we expected this to exceed $14 USD per hour based on a generous (conservatively high) estimate of completion time. All annotators reported their work time was less than this estimate.

notators found difficult or ambiguous. For each sentence presented to annotators, we ask “If you found this example difficult or ambiguous please explain why” and ask them to provide a short written response in a provided text-box. This follows recent work that that has emphasized the importance of annotator disagreement not necessarily always as error in annotation but instead as ambiguity that is inherent to natural language and a potential useful signal for downstream analyses (Dumitrache et al., 2018; Paun et al., 2018; Pavlick and Kwiatkowski, 2019; Keith et al., 2020).

Annotators remarked on several types of text they were uncertain about: agents of actions who were not explicitly mentioned but implicitly police, named entities whose status as police is ambiguous, confusion about what precisely constitutes an “arrest”, and confusion arising from the lack of specific cultural knowledge (e.g., around the Indian crowd-control tactic of “lathi charging”). In the appendix, see Table A3 for examples and Table A2 for a categorization of free text responses.

5 Baseline Models

We test several baseline models, all requiring no annotation (and thus most realistic for the social science use case), and assess their performance on INDIAPOLICEEVENTS.

Keyword matching. Boolean keyword queries are a very common social science approach to document classification (e.g. Nielsen (2013); Chadefaux (2014); D’Orazio et al. (2014); Baker et al. (2016)), since they are simple, transparent, and widely supported in user software. We use conjunctive normal form rules, where inferring an event class for a sentence requires matching any term from a *police* keyword list (including both common nouns and names of major police and security institutions), as well as an event keyword. To construct the keyword lists, a domain expert coauthor first manually generates a list of seed keywords for the semantic categories *police*, *kill*, *arrest*, *intervention*, and *force*. To address lexical coverage, we then expand the keywords through word2vec (Mikolov et al., 2013) nearest neighbors, filtered to semantically equivalent words by the domain expert.[11] This process is repeated using WordNet synonym sets for lookup (Miller, 1995), resulting in 217 keywords total; see appendix (§C.2) for details.

RoBERTa+MNLI. Given two input sentences, a premise and hypothesis, the task of natural language inference (NLI) is to predict whether the premise *entails* or *contradicts* the hypothesis or does neither (*neutral*) (Bowman et al., 2015; Williams et al., 2018). Previous work has shown promise of NLI transfer learning for events: Sarwar and Allan (2020) show Sentence-BERT embeddings (Reimers et al., 2019) learned from NLI data are effective on ACE-like event retrieval; Clark et al. (2019) find for *BoolQ*, their dataset of naturally occurring boolean questions, that transfer learning from NLI data is more effective than transferring from QA or paraphrase data. We follow Clark et al. (2019)’s example and use “off-the-shelf” RoBERTa (Liu et al., 2019) fine-tuned on the MNLI corpus (Williams et al., 2018). The model takes a sentence and a declarative form of an event class question as input (§C.1), and we use its predicted probability of entailment as the probability of the event class. For document ranking, we create a document score by taking the maximum predicted probability over sentences. Future experiments could vary the amount of text (sentence vs. passage vs. document) used as input to the model.

BM25+RM3. Weighted term matching between a query and document is a strong competitor to neural ranking methods (Craswell et al., 2020; Lin, 2019), via, for example, BM25 scoring with RM3 query expansion (Lavrenko and Croft, 2001). With the Anserini BM25 implementation (Yang et al., 2018a), we set $k_1 = 0.9$ and $b = 0.4$, and conduct RM3 expansion of the query to terms found in the top $k = 10$ BM25-retrieved documents, following Lin (2019)’s hyperparameter settings. As the input query, this set of models uses the natural language questions described in §4.2. Appendix Table A5 contains full results.

ELECTRA+MS MARCO. Fine-tuned BERT (Devlin et al., 2019) and other large-scale language models have been used extensively for document ranking in information retrieval (IR) (Zhan et al., 2020; Zhang et al., 2020; Dai and Callan, 2019; MacAvaney et al., 2019). We use a competitive off-the-shelf model that uses the ELECTRA variant of BERT (Clark et al., 2020) fine-tuned on MS MARCO (Reimers et al., 2019). MS MARCO is a large-scale reading comprehension dataset in which questions are sampled from anonymized web

[11] We train word2vec on every article in the *Times of India* from 2002 (the same corpus as our dataset, 69,000 articles) plus another 100,000 articles from *The Hindu*, another English-language newspaper in India. We inspect each keyword’s 20 nearest neighbors with highest cosine similarity.

queries, and answers to the queries are generated by crowdworkers (Nguyen et al., 2016). It was used in the 2019 TREC Deep Learning track on document and passage retrieval (Craswell et al., 2020). We use Reimers et al. (2019)'s pretrained ELECTRA+MS MARCO model. Inputs are INDIAPOLICEEVENTS passages consisting of three-sentence sliding windows with stride of one sentence and queries are the event class questions described in §4.2. Following related work (Zhan et al., 2020), we take the maximum score of all passages as the document score.

6 Results

We report the performance of the baseline models (§5) on our three tasks in Table 2 and in Figure 3.

6.1 Task 1: Sentence classification.

For sentence classification,[12] Table 2 shows that the keyword matching method slightly outperforms RoBERTa+MNLI on F1 for ANY ACTION and FORCE, which we suspect is due to the keyword method having better access to synonyms of "police" (e.g. "jawan", "RPF") particular to the *Times of India* via its *word2vec* expansion. However, RoBERTa+MNLI achieves a higher F1 score on KILL, ARREST, and FAIL TO ACT. We need further controlled experiments to understand how the concreteness of the event class, importance of identifying events' agents, and formulation of the query (e.g. "Did police use force or violence?" vs. "Were police violent?") affect the results of contextualized language models. Table 2 also shows poor performance of our keyword matching method on FAIL TO ACT (F1=0.05); however, a large-scale contextual language model seems to be able to better distinguish the semantics of the event class (F1=0.48). The Task 1 plot in Figure 3 shows that across all labels, RoBERTa+MNLI has higher recall than the keyword method for every event class. If social scientists plan to use these sentence classification methods in a semi-automated fashion (as we suggest in §1), selecting models like RoBERTa+MNLI that achieve higher recall may be important.

6.2 Task 2: Document ranking.

For Task 2, we report average precision and a new metric—the proportion of documents that would have to be read to achieve recall equal to X (*PropRead@RecallX*). We use $X = 0.95$ because social scientists typically use 95% cutoffs for significance and sampling error.[13] We leave to future work estimating recall on a corpus without ground truth. Table 2 shows that RoBERTa+MNLI outperforms both BM25 and ELECTRA+MS MARCO on both average precision and PropRead@Recall95 across all event classes. We hypothesize this is because natural language inference is a task that is much more aligned with the semantic-oriented precision at which we want to rank documents. In contrast, the MS MARCO dataset is constructed for a much higher level information need, and documents that are "relevant" could potentially not entail the semantic event class of interest. As Figure 3 shows, if a social scientist was presented with a ranked list of documents from RoBERTa+MNLI, they would only have to read 5% of the entire corpus to achieve 95% recall on KILL. RoBERTa+MNLI also does well on ARREST and FORCE with 0.17 and 0.20 PropRead@Recall95 respectively. There is consistently more difficulty across all models for ANY ACTION and FAIL TO ACT. We speculate this is because ANY ACTION is the class with the greatest prevalence, and thus is more difficult to achieve higher recall.

6.3 Task 3: Temporal Aggregates.

Figure 4 compares the outputs of three systems on FAIL TO ACT: gold-standard human annotations, keyword matching, and RoBERTa+MNLI. For this event class, both automated methods under-count the number of events marked by human annotators. In contrast, the automated techniques tend to *over*count other event types (see Appendix, Figure A5 for plots of the other event classes). While the the overall temporal trend is broadly consistent across the three methods, the decreased accuracy of the automated methods could lead to attenuation bias if they were used as input to statistical models. A qualitative examination of the extracted events also reveals the need for future work in temporal linking models: most of the events after March 25

[12] We do not evaluate sentences with less than 5 tokens as many of these sentences are due to sentence segmentation errors. After this filtering, the number of remaining sentences we evaluate on is 18,645.

[13] We note that 5% recall error is not equivalent to a 5% sampling error. In practice, researchers are more concerned with whether data is missing at random.

	Task 1: Sent. Cls.		*Task 2: Document Ranking*						*Task 3: Temp. Aggs.*	
	Keyw.	R+MNLI	BM25		E+MSM		R+MNLI		Keyw.	R+MNLI
Event Class	F1 ↑	F1 ↑	AP ↑	PR ↓	AP ↑	PR ↓	AP ↑	PR ↓	ρ ↑	ρ ↑
KILL	0.50	**0.74**	0.30	0.29	0.65	0.27	**0.96**	**0.05**	0.70	**0.78**
ARREST	0.48	**0.62**	0.68	0.36	0.72	0.67	**0.91**	**0.17**	0.71	**0.85**
FAIL TO ACT	0.05	**0.48**	0.27	0.77	0.36	0.87	**0.63**	**0.76**	0.42	**0.60**
FORCE	**0.65**	0.62	0.24	0.43	0.64	0.45	**0.90**	**0.20**	**0.89**	0.86
ANY ACTION	**0.67**	0.57	0.53	0.85	0.83	0.88	**0.89**	**0.62**	0.86	**0.90**

Table 2: Evaluation of two classification methods (Keyw., R+MNLI) and three ranking models (BM25, E+MSM, R+MNLI) for IndiaPoliceEvents's three tasks. Bolded numbers indicate the model that performs best on each metric and event class. **Task 1** evaluates sentence-level F1 for sentence-level keyword matching (Keyw.) and RoBERTa fine-tuned on MNLI (R+MNLI) (Liu et al., 2019). **Task 2** evaluates average precision (AP) and proportion of the corpus needed to be read in order to achieve 95% recall (PR, or PropRead@Recall95) for ranking models BM25 (Yang et al., 2018a), off-the-shelf ELECTRA language model fine-tuned on MS MARCO (E+MSM; (Reimers et al., 2019)), as well as R+MNLI's probabilistic output. **Task 3** evaluates Spearman's rank correlation coefficient (ρ) between predicted versus gold-standard counts of documents with the relevant event, for each day in March 2002. For each metric, we indicate whether a higher (↑) or lower (↓) score is better.

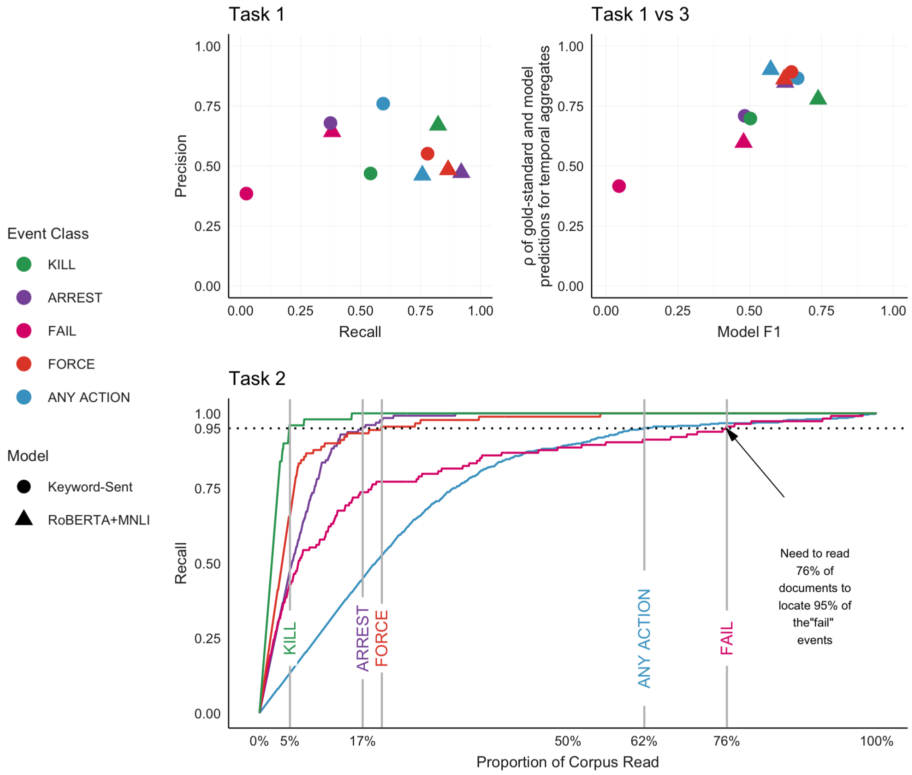


Figure 3: Keyword and RoBERTA+MNLI performance on three metrics. **(Task 1)** Precision and recall at the sentence level for two models on each semantic event class. **(Task 1 vs 3)** Sentence-level model F1 scores (x-axis) versus Spearman's ρ of hand-annotated gold-standard and model predictions for temporal aggregates (y-axis). **(Task 2)** For each class under RoBERTa MNLI, the gain curves (Grossman et al., 2016), marking PropRead@Recall95: What percentage of the ranked corpus would a researcher need to read in order to find 95% of each event classes' mentions?

are describing events from earlier in March that were being reported in the context of investigations into the violence. Table 2 shows that for all event classes except for FORCE RoBERTa+MNLI has a

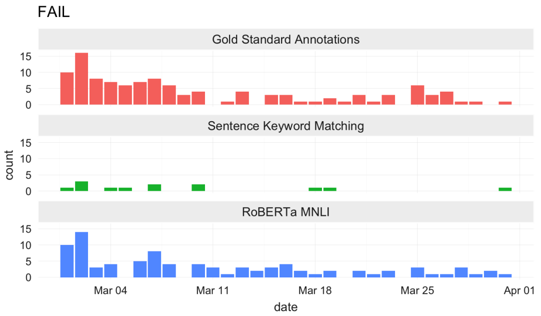


Figure 4: Number of documents per day containing a "police failing to act/standing by" event, comparing outputs from **(Top)** human annotations **(Middle)** keyword matching (Spearman's $\rho = 0.51$, comparing with gold standard) and **(Bottom)** RoBERTa finetuned on MNLI ($\rho = 0.60$).

higher Spearman's ρ[14] between the predicted versus gold-standard document counts. The Task 1 vs. 3 plot in Figure 3 shows an approximately linear relationship between the F1 scores of sentence-level models and Spearman's ρ, suggesting there is promise that NLP research focused on sentence-level models could be of use to social scientists who care about corpus-level evaluation.

6.4 Qualitative error analysis.

We manually analyze the false positives and false negatives of our best-performing baseline model, RoBERTa+MNLI. Some false positives are due to lexical semantic misunderstandings: the model often mistakes "shot" for KILL, and assigns high probability to negative FORCE sentences such as, *"The police escorting the vehicle fired into the air and dispersed the mob."* The model also has difficulty identifying the police as agents: for example, it assigns high probability to the negative KILL sentences: *"[. . .] scores of people have been killed in rural Gujarat due to police failure [. . .],"* and *"Police said that two persons had been killed in Vijaynagar [. . .]"*. Other errors are due to *hypotheticals*: the model assigns high probability to the negative KILL sentences *"He alleged BJP's hand in the murder of [. . .]"* and *"Achar claims she was an eye-witness to police complicity in the violence."* Many of the model's false negatives are due to necessary multi-sentence context (which RoBERTa+MNLI does not have as it only takes single sentences as input). For instance, the model assigns low probability to the positive KILL sentence *"Four persons have been killed and five are injured."* and FORCE sentence *"One person was injured and rushed to the SSG hospital"*; if one reads the proceeding context of both of these sentences it is clear that police are the agents of the actions.

7 Discussion and Future Work

The dataset, tasks, and evaluations we present in this work are driven by the needs of social scientists: we assess the performance of zero-shot models on metrics important to applied researchers, including recall against a fully annotated corpus and performance at temporally aggregated levels. We find cause for optimism for social scientists using BERT-style pre-trained models on their tasks. These models could potentially be used in place of social scientists' existing keyword-based classifiers, although we caution accuracy is far from perfect and applied researchers will need to extensively validate model outputs. Even with imperfect classification accuracy, we believe these zero-shot models show promise for decreasing human annotation effort by reducing the proportion of the corpus read to achieve a specific recall level (the metric we call PropRead@RecallX).

Future work can extend our dataset creation process to new semantic event classes, such as protests, communal violence itself, and other forms of participation in political and social activity. Additional annotated datasets could allow researchers to generalize the performance of zero-shot language models to new domains and event classes. Finally, tasks such as temporal and geographic linking, event deduplication and coreference, and identifying hypothetical events are unsolved but are major obstacles for applied social scientists working with automatically extracted events.

8 Acknowledgments

For helpful comments, we're grateful to Aidan Milliff, the UMass NLP group, and anonymous ACL reviewers. This work was funded by a Kaggle Open Data Science grant, and additionally: *AH:* a National Science Foundation Graduate Research Fellowship; *KK:* a Bloomberg Data Science PhD Fellowship; *SS:* the Center for Intelligent Information Retrieval (CIIR); *BO:* National Science Foundation IIS-1845576 and IIS-1814955.

[14] See Table A6 for mean absolute error scores.

9 Ethical Considerations and Broader Impact

To ensure the replicability of our work and to further research into event extraction systems for social science research, we are making the text of the news articles available to researchers alongside our annotations. While all articles were obtained from a public website without login credentials, the applicability of copyright restrictions is relevant to address.

We believe the research benefits and the limited harms to the copyright holders justify this use, due to the four criteria considered in the fair use doctrine in U.S. copyright law (U.S. Copyright Office, 2021): (1) the non-commercial, nonprofit educational purpose of our use of the text, (2) the factual nature of the news reports, (3) the limited substitutability of our dataset for the original news site,[15] and (4) our expectation that our limited corpus will not harm the market for readers of the news site.

The issue of copyright status within NLP-oriented corpora is of increasing interest. Sag (2019) argues machine learning uses of text is non-expressive and therefore falls under fair use, and Geiger et al. (2018) study the issue in the context of proposals for E.U. law. Bandy and Vincent (2021) investigate BooksCorpus, a previously poorly documented corpus widely used for training language models, finding it contains large amounts of copyrighted work, highlighting how current data curation practices in machine learning (and adjacent) communities need improvement (Paullada et al., 2020; Jo and Gebru, 2020).

We also acknowledge the sensitivities around this period of violence in India. Its significance motivates computational work to enable more effective study of it and related episodes, but our news-derived data on its own, in the absence of deeper qualitative work, does not permit us to draw new substantive conclusions about the causes and consequences of the violence in Gujarat in 2002. We defer to the large scholarly and journalistic literature on the violence; see references in §1 and §4.

[15] We do not republish the texts as consumer-accessible webpages, but instead are only contained within a JSON structured format.

References

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89, Vancouver, Canada. Association for Computational Linguistics.

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53.

Avi Arampatzis, Jaap Kamps, and Stephen Robertson. 2009. Where to stop reading a ranked list? threshold optimization using truncated score distributions. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 524–531.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Scott R Baker, Nicholas Bloom, and Steven J Davis. 2016. Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636.

Jack Bandy and Nicholas Vincent. 2021. Addressing "documentation debt" in machine learning research: A retrospective datasheet for BookCorpus. *arXiv preprint arXiv:2105.05241*.

Amrita Basu. 2015. *Violent conjunctures in democratic India*. Cambridge University Press.

John Beieler. 2016. Generating politically-relevant event data. *CoRR*.

John Beieler, Patrick T Brandt, Andrew Halterman, Erin Simpson, and Philip A Schrodt. 2016. Generating political event data in near real time: Opportunities and challenges. In R. Michael Alvarez, editor, *Computational Social Science*, chapter 98. Cambridge University Press.

Ward Berenschot. 2012. *Riot politics: Hindu-Muslim violence and the Indian state*. Rupa Publications.

Elizabeth Boschee. 2016. Solutions for coding societal events. Technical report, Raytheon BBN Technologies Corp. Cambridge United States.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Stephen M Shellman, James Starz, and Michael D Ward. 2015. ICEWS coded event data. In *Harvard Dataverse, V9, http://dx.doi.org/10.7910/DVN/28075*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Robert Brathwaite and Baekkwan Park. 2018. Measurement and conceptual approaches to religious violence: The use of natural language processing to generate religious violence event-data. *Politics and Religion*, pages 1–42.

Thomas Chadefaux. 2014. Early warning signals for war in the news. *Journal of Peace Research*, 51(1):5–18.

Erica Chenoweth and Orion A Lewis. 2013. Unpacking nonviolent campaigns introducing the NAVCO 2.0 dataset. *Journal of Peace Research*, 50(3):415–423.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ICLR*.

Jim Cowie and Wendy Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 Deep Learning track. *arXiv preprint arXiv:2003.07820*.

Zhuyun Dai and J. Callan. 2019. Deeper text understanding for ir with contextual neural language modeling. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ursula Daxecker, Elio Amicarelli, and Alexander Jung. 2019. Electoral Contention and Violence (ECAV): A new dataset. *Journal of Peace Research*, 56(5):714–723.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Raheel Dhattiwala and Michael Biggs. 2012. The political logic of ethnic violence: The anti-Muslim pogrom in Gujarat, 2002. *Politics & Society*, 40(4):483–516.

Erika Doggett and Alejandro Cantarero. 2016. Identifying eyewitness news-worthy events on Twitter. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 7–13.

Vito D'Orazio, Steven T Landis, Glenn Palmer, and Philip Schrodt. 2014. Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political analysis*, 22(2):224–242.

Ellen A Drost et al. 2011. Validity and reliability in social science research. *Education Research and perspectives*, 38(1):105.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing ground truth for medical relation extraction. *ACM Transactions on Interactive Intelligent Systems*, 8(2):1–20.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke S. Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *ACL*.

Christophe Geiger, Giancarlo Frosio, and Oleksandr Bulayenko. 2018. The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market—Legal Aspects. Policy Department for Citizens' Rights and Constitutional Affairs, Directorate General for Internal Policies of the Union, European Parliament, PE 604.941, February 2018. https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf accessed 2021-05-25.

Deborah J. Gerner, Philip A Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. 2002. Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. *International Studies Association, New Orleans*.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In *International summer school on information extraction*, pages 10–27. Springer.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

M. R. Grossman, G. Cormack, and Adam Roegiest. 2016. TREC 2016 Total Recall Track overview. In *TREC*.

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

Alex Hanna. 2014. Developing a system for the automated coding of protest event data. *Available at SSRN: http://ssrn.com/abstract=2425232*.

Alex Hanna. 2017. MPEDS: Automating the generation of protest event data. *SocArXiv https://osf.io/preprints/socarxiv/xuqmv*.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R. Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–268, Berlin, Germany. Association for Computational Linguistics.

Human Rights Watch. 2002. "We have no orders to save you": State participation and complicity in communal violence in Gujarat. *Human Rights Watch Report*, 14(3).

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Katherine Keith, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor. 2017. Identifying civilians killed by police with distantly supervised entity-event extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1547–1557.

Katherine Keith, Christoph Teichmann, Brendan O'Connor, and Edgar Meij. 2020. Uncertainty over Uncertainty: Investigating the Assumptions, Annotations, and Text Measurements of Economic Policy Uncertainty. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 116–131.

Gary King. 1989. *Unifying political methodology: The likelihood theory of statistical inference*. Cambridge University Press.

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(03):617–642.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 120–127.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *CoNLL 2017*, page 333.

Yan Liang, Khaled Jabr, Christan Grant, Jill Irvine, and Andrew Halterman. 2018. New techniques for coding political events across languages. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 88–93. IEEE.

Yen-Chieh Lien, Daniel Cohen, and W Bruce Croft. 2019. An assumption-free approach to the dynamic truncation of ranked lists. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 79–82.

Jimmy Lin. 2019. The simplest thing that can possibly work: Pseudo-relevance feedback using text classification. *ArXiv*, abs/1904.08861.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2020. Sparse, dense, and attentional representations for text retrieval. *arXiv preprint arXiv:2005.00181*.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1101–1104.

Andrew McCallum. 2005. Information extraction: Distilling structured data from unstructured text. *Queue*, 3(9):48–57.

Graham McDonald, Craig Macdonald, and Iadh Ounis. 2018. Active learning strategies for technology assisted sensitivity review. In *European Conference on Information Retrieval*, pages 439–453. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*, to appear.

Hannes Mueller and Christopher Rauh. 2017. Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, pages 1–18.

Minh Nguyen and Thien Huu Nguyen. 2018. Who is killed by police: Introducing supervised attention for hierarchical LSTMs. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2277–2287, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Richard A Nielsen. 2013. Rewarding human rights? Selective aid sanctions against repressive states. *International Studies Quarterly*, 57(4):791–803.

Brendan O'Connor, Brandon Stewart, and Noah A Smith. 2013. Learning to extract international relations from political context. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1.

Javier Osorio and Alejandro Reyes. 2017. Supervised event coding from text written in Spanish: Introducing Eventus ID. *Social Science Computer Review*, 35(3):406–416.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*. NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-analyses (ML-RSA), Virtual.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Vinodkumar Prabhakaran, Camilla Griffiths, Hang Su, Prateek Verma, Nelson Morgan, Jennifer L Eberhardt, and Dan Jurafsky. 2018. Detecting institutional dialog acts in police traffic stops. *Transactions of the Association for Computational Linguistics*, 6:467–481.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Benjamin James Radford. 2016. *Automated learning of event coding dictionaries for novel domains with an application to cyberspace*. Ph.D. thesis, Duke University.

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Matthew Sag. 2019. The new legal landscape for text mining and machine learning. *Journal of the Copyright Society of the USA*, 66:1–64. Accessed at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606.

Sheikh Muhammad Sarwar and James Allan. 2019. Searchie: A retrieval approach for information extraction. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 249–252.

Sheikh Muhammad Sarwar and James Allan. 2020. Query by example for cross-lingual event retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1601–1604.

Philip A Schrodt, Shannon G Davis, and Judith L Weddle. 1994. Political science: KEDS—a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561–587.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Gabriel Stanovsky, Meni Adler, and Ido Dagan. 2016. Specifying and annotating reduced argument span via QA-SRL. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Kadayam Suryanarayanan Subramanian. 2007. *Political violence and the police in India*. SAGE Publications India.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

U.S. Copyright Office. 2021. More information on fair use. https://www.copyright.gov/fair-use/more-info.html http://web.archive.org/web/20210521070009/https://www.copyright.gov/fair-use/more-info.html.

Siddharth Varadarajan. 2002. *Gujarat, the Making of a Tragedy*. Penguin Books India.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.

Hanna Wallach. 2018. Computational social science$\neq$ computer science+ social data. *Communications of the ACM*, 61(3):42–44.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019. Evidence sentence extraction for machine reading comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 696–707.

Steven I Wilkinson. 2006. *Votes and violence: Electoral competition and ethnic riots in India*. Cambridge University Press.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018a. Anserini: Reproducible ranking baselines using lucene. *ACM J. Data Inf. Qual.*, 10:16:1–16:20.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018b. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized Text Embeddings for First-Stage Retrieval. *arXiv preprint arXiv:2006.15498*.

Xinyu Zhang, Andrew Yates, and Jimmy Lin. 2020. A little bit is worse than none: Ranking with limited training data. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 107–112, Online. Association for Computational Linguistics.

Appendix

A Annotation Details

We provide details on our annotation process here, including the semantic event class definitions we provided to annotators, the per-class agreement statistics, statistics on the time it took to annotate, and further qualitative analysis of the annotations.

All results reported in this appendix correspond to responses to the annotation questions (shown in Table A1), which are slightly different from the main semantic event classes reported in the main paper, as described in Table A1's caption and Footnote 9.

A.1 Annotator Instructions

To train our annotators, we provided them with a the question for each semantic event class, a short description to clarify the question, and an example positive sentence (Figure A1). We met with the annotators as a group to talk through the document and then gave them a training round with documents we had previously annotated. Based on that training round, we added frequently asked questions to the instructions document, provided individual feedback to annotators, and then began the production annotation process on our corpus.

A.2 Annotation Interface

Figure A2 shows a stylized version of the custom interface we built using the Prodigy annotation tool (Montani and Honnibal, 2018). Annotators are presented with an entire document, with sentences sequentially highlighted. For each highlighted question, they are asked each of the questions. If the sentence contains a positive answer to the question(s), they select the corresponding box(es) and advance to the next sentence.

A.3 Multi-sentence labels

We record whether annotators report using information from other sentences in the document to annotate the current sentence. Specifically, we provide a checkbox in the interface with the label "I used information from other sentences to answer the question". We collected this information in order to understand the number of sentences that could be classified on their own and how many needed broader document context. We caution that we left the interpretation of the sentence up to each annotator and did not train them or compare their usage of this label as we did with other labels. We do not use the labels in our analysis but provide them in our dataset to potentially help future research.

A.4 Annotator Agreements

We calculated inter-annotator agreement, both raw agreement and Krippendorff's alpha, for all annotators on our corpus (Table A1). Because the event classes are rare in our corpus, we prefer Krippendorff's alpha over raw agreement, which is inflated by the large number of zeros in our data. After half of the documents were annotated, we calculated agreement to check for annotators with high disagreement. We found one annotator with high disagreement on the KILL class and provided updated instructions for them. We used the final agreement rates to select the three annotators with the highest agreement rates with the full set of annotators to serve as our adjudicators in the final round of annotations.

A.5 Annotation Timing

Figure A3 shows the distribution of the time that annotators took to annotate each document.

B Properties of Annotated Data

B.1 Event locality

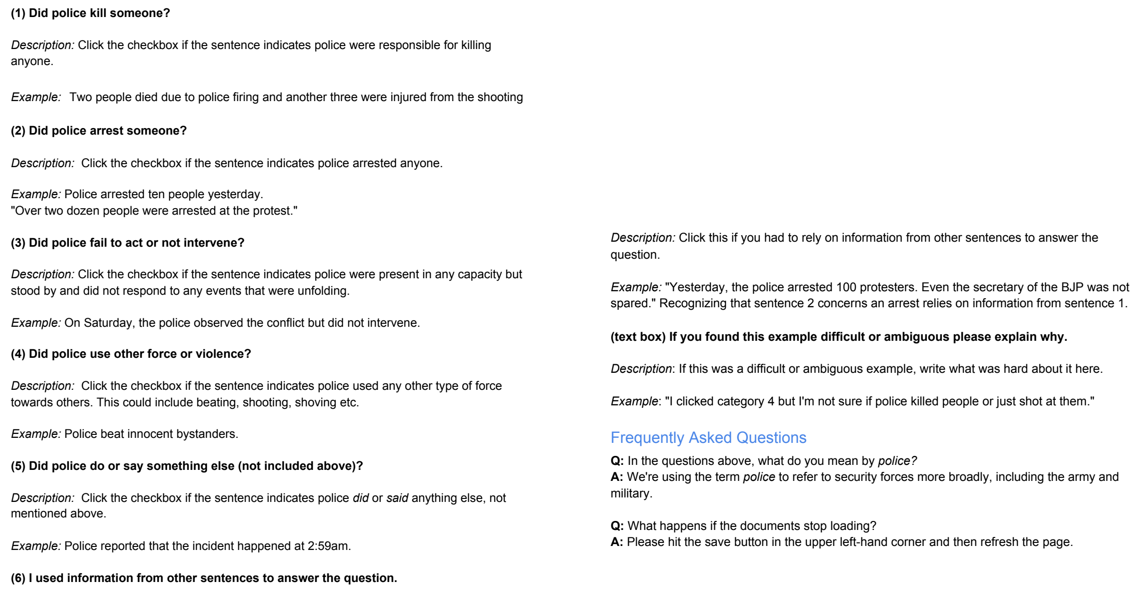
Figure A4 shows the the label density for each event class in different sections in a document. To compute the label density, we partition the sentences in each document into ten equal and ordered sections, where the first section indicates the earliest location in a document. Then we compute the number of positive event in those sections. Figure A4 shows that except for ARREST, label density is not high in the initial sections of a document. In information retrieval and news summarization, typically the first k tokens are assumed a good approximation for document representation (Dai and Callan, 2019), which our dataset seems to present contradictory evidence.

B.2 Analysis of Free-Text Explanations

We analyzed the free text explanations given by annotators, grouping them into non-exclusive categorizes. The most common categories of annotator explanations are shown in Table A2.

B.3 Interesting Examples

Table A3 shows a selection of sentences from our corpus that illustrate challenging annotation deci-



(1) Did police kill someone?

Description: Click the checkbox if the sentence indicates police were responsible for killing anyone.

Example: Two people died due to police firing and another three were injured from the shooting

(2) Did police arrest someone?

Description: Click the checkbox if the sentence indicates police arrested anyone.

Example: Police arrested ten people yesterday.
"Over two dozen people were arrested at the protest."

(3) Did police fail to act or not intervene?

Description: Click the checkbox if the sentence indicates police were present in any capacity but stood by and did not respond to any events that were unfolding.

Example: On Saturday, the police observed the conflict but did not intervene.

(4) Did police use other force or violence?

Description: Click the checkbox if the sentence indicates police used any other type of force towards others. This could include beating, shooting, shoving etc.

Example: Police beat innocent bystanders.

(5) Did police do or say something else (not included above)?

Description: Click the checkbox if the sentence indicates police *did* or *said* anything else, not mentioned above.

Example: Police reported that the incident happened at 2:59am.

(6) I used information from other sentences to answer the question.

Description: Click this if you had to rely on information from other sentences to answer the question.

Example: "Yesterday, the police arrested 100 protesters. Even the secretary of the BJP was not spared." Recognizing that sentence 2 concerns an arrest relies on information from sentence 1.

(text box) If you found this example difficult or ambiguous please explain why.

Description: If this was a difficult or ambiguous example, write what was hard about it here.

Example: "I clicked category 4 but I'm not sure if police killed people or just shot at them."

Frequently Asked Questions

Q: In the questions above, what do you mean by *police?*
A: We're using the term *police* to refer to security forces more broadly, including the army and military.

Q: What happens if the documents stop loading?
A: Please hit the save button in the upper left-hand corner and then refresh the page.

Figure A1: Instructions provided to annotators providing additional guidance on how to interpret the questions, giving an example positive sentence, and clarifying other issues that arose in training.

Question	Agr. (all)	Agr. (1+)	Krip. (all)	Krip. (1+)	Support (1+)
(1) “Did police kill someone?”	0.998	0.984	0.753	0.751	108
(2) “Did police arrest someone?”	0.993	0.949	0.734	0.710	328
(3) “Did police use other force or violence?”	0.995	0.961	0.704	0.686	227
(4) “Did police fail to act or not intervene?”	0.988	0.907	0.418	0.377	339
(5) “Did police say or do anything else?”	0.942	0.555	0.587	0.086	2142

Table A1: Sentence-level agreement, Krippendorff, and support for each question answered by annotators. “All” refers to all 20,527 annotated sentences in the corpus and “(1+)” refers to a subset of the corpus that excludes sentences that both annotators agree do not have police actions. Questions 1, 2, and 4 map to KILL, ARREST, and FAIL TO ACT, respectively; FORCE is defined as (*1 OR 3*), and ANY ACTION is (*1 OR 2 OR 3 OR 5*), as described in Footnote 9.

sions for our annotators or interesting ambiguity in the sentences.

C Modeling Details

C.1 Declarative versions of questions

We use the following declarative versions of event class labels as input to RoBERTa+MNLI:

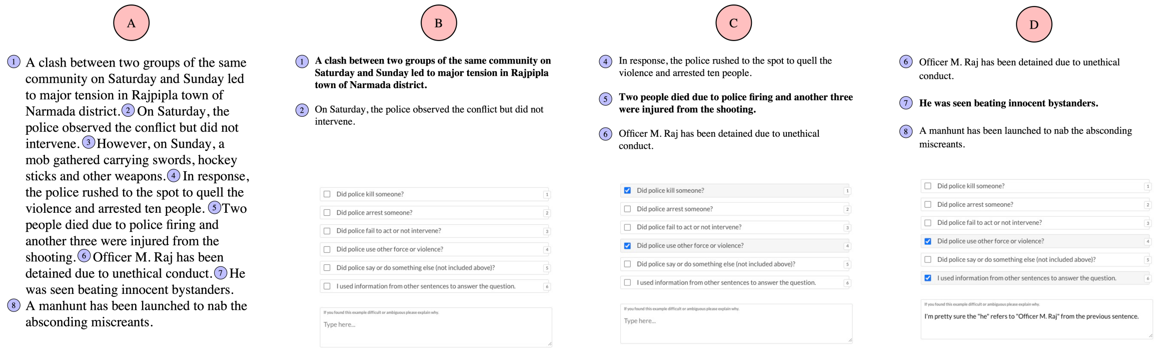
- KILL: “Police killed someone.”
- ARREST: “Police arrested someone.”
- FAIL TO ACT: “Police failed to intervene.”
- FORCE: “Police used violence.”
- ANY ACTION: “Police did something.”

C.2 Keyword Approach

We report here the terms used in the keyword matching. These terms were generated using subject matter expertise and expanded using WordNet and a custom word2vec model trained on the complete set of *Times of India* articles from 2002 and 100,000 additional articles from the Indian newspaper *The Hindu*. The expanded set was filtered using subject matter expertise. We report the keywords on the following categories:

Police: police, policemen, cop, cops, constables, constables, jawan, jawans, grp , cid , rpf , stf , bsf , dcp , dsp , ssp , sho , cisf , dgp.

Kill: kill, kills, killed, killing, lynch, lynched, lynching, annihilate, annihilating, annihilated, annihilates, drown, drowning, drowned, drowns, massacre, massacring, massacred, massacres, slaughter, slaughtering, slaughtered, slaughterers,



1 A clash between two groups of the same community on Saturday and Sunday led to major tension in Rajpipla town of Narmada district. 2 On Saturday, the police observed the conflict but did not intervene. 3 However, on Sunday, a mob gathered carrying swords, hockey sticks and other weapons. 4 In response, the police rushed to the spot to quell the violence and arrested ten people. 5 Two people died due to police firing and another three were injured from the shooting. 6 Officer M. Raj has been detained due to unethical conduct. 7 He was seen beating innocent bystanders.
8 A manhunt has been launched to nab the absconding miscreants.

1 **A clash between two groups of the same community on Saturday and Sunday led to major tension in Rajpipla town of Narmada district.**

2 On Saturday, the police observed the conflict but did not intervene.

- □ Did police kill someone? 1
- □ Did police arrest someone? 2
- □ Did police fail to act or not intervene? 3
- □ Did police use other force or violence? 4
- □ Did police say or do something else (not included above)? 5
- □ I used information from other sentences to answer the question. 6

If you found this example difficult or ambiguous please explain why.
Type here...

4 In response, the police rushed to the spot to quell the violence and arrested ten people.

5 **Two people died due to police firing and another three were injured from the shooting.**

6 Officer M. Raj has been detained due to unethical conduct.

- ☑ Did police kill someone? 1
- □ Did police arrest someone? 2
- □ Did police fail to act or not intervene? 3
- ☑ Did police use other force or violence? 4
- □ Did police say or do something else (not included above)? 5
- □ I used information from other sentences to answer the question. 6

If you found this example difficult or ambiguous please explain why.
Type here...

6 Officer M. Raj has been detained due to unethical conduct.

7 **He was seen beating innocent bystanders.**

8 A manhunt has been launched to nab the absconding miscreants.

- □ Did police kill someone? 1
- □ Did police arrest someone? 2
- □ Did police fail to act or not intervene? 3
- ☑ Did police use other force or violence? 4
- □ Did police say or do something else (not included above)? 5
- ☑ I used information from other sentences to answer the question. 6

If you found this example difficult or ambiguous please explain why.
I'm pretty sure the "he" refers to "Officer M. Raj" from the previous sentence.

Figure A2: Illustration of the dataset annotation interface given an example document. In practice, annotators view and label *all* sentences in the document, but this figure highlights *three* informative sentence examples. (A) An example document with post-hoc numbering of sentences. (B) The user sees a bold sentence its context and then is asked a series of yes/no questions about the bold sentence. For this example, the annotators do not check any boxes (the answer to all the questions is *no*). (C) For this bold sentence, annotators check the boxes (*yes* answers) for the questions "Did police kill someone?" and "Did police use other force or violence." (D) For this bold sentence, annotators check the box for "Did police use other force or violence?", select "I used information from other sentences to answer the question," and provide a free-text explanation for why they thought the example was difficult.

Author-assigned category	**Count**
No explicit agent	63
Agent may not be police	57
Police are mentioned but not agents	37
Hypothetical or future events	35
Failing to act vs. acting and failing	30
True ambiguity in language	27
Ambiguity in "arrest"	26
Total free text explanations	311
Total sentences with explanations	299
Total sentences with police activity*	2,783
Total sentences in corpus	21,391

Table A2: Top categories of free-text explanations by annotators. Text explanations can be assigned to multiple categories. *Here, "total sentences with police activity" means at least one annotator noted it had police activity.

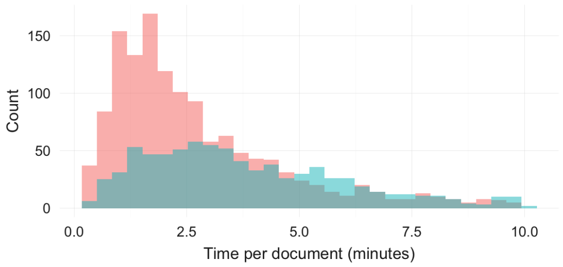


Figure A3: Time to annotate a document in minutes. The median time to annotate one document is 3.0 minutes, with documents that contain no police events (red) taking less time than documents with police activity (blue). Not shown are 366 out of 2,514 documents taking longer than 10 minutes. Median document length is 15 sentences, min length=2, 25% = 9, 75%=22, max=98.

butcher, butchering, butchered, butchers, poison, poisoning, poisoned, poisons, exterminate, exterminating, exterminated, exterminates, strangle, strangling, strangled, strangles, impale, impaling, impaled, impales, murder, murdering, murdered, murders, execute, executing, executed, executes.

Arrest: arrest, arresting, arrested, arrests, nab, nabbing, nabbed, nabs, detain, detaining, detained, detains, book, booking, booked, books, chargesheet, chargesheeting, chargesheeted, chargesheets, apprehend, apprehending, apprehended, apprehends, seize, seizing, seized, seizes, collar, collaring, collared, collars.

Sentence from INDIAPOLICEEVENTS	Author comment
At one point, the men in khaki seemed to have outnumbered the vhp workers	Keyword matching would have missed "men in khaki" as a reference to police.
The police have rounded up 1,740 VHP activists headed for Ayodhya.	Annotators flagged ambiguity in "rounded up" vs. "arrested"
One of them who was on duty on December 16 even recollected how he and another colleague had to burst the tear gas shell themselves as the constables deliberately looked the other way.	Annotators flagged this sentence as two separate police agencies acted and failed to act.
Police on Friday lathicharged ram sevaks who attempted to rush towards the make-shift temple in the disputed site here giving some anxious moments to security forces.	"Lathi charges" are an Indian riot control tactic that our United States-based annotators were not familiar with.
Meanwhile the district administration has tightened the security in and around the temple city.	Many annotators flagged sentences where police are implicitly the agents.

Table A3: Example sentences illustrating several of the challenges of annotating the documents or in applying existing models. We provide our own commentary on why the sentences are difficult.

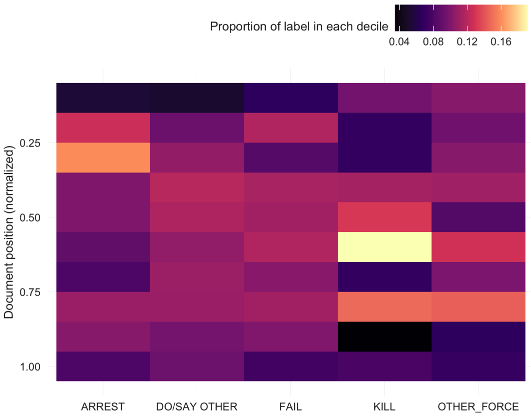


Figure A4: The location within a document of answers to each of the questions. Questions are often answered in the second half of the document.

Intervention: intervene, intervening, intervened, intervenes, intervention, interfere, interfering, interfered, interferes, stand by, standing by, stood by, stands by, abstain, abstaining, abstained, abstains.

Force: fire, firing, fired, fires, stone-pelt, stone-pelting, stone-pelted, stone-pelts, pelt stones, pelting stones, pelts stones, pelted stones, beat, beating, beaten, beats, whip, whipping, whipped, whips, bash, bashing, bashed, bashes, choke, choking, choked, chokes, wound, wounding, wounded, wounds, strong-arm, strong-arming, strong-armed, strong-arms, pistol-whip, pistol-whipping, pistol-whipped, pistol-whips, lash, lashing, lashed, lashes, trounce, trouncing, trounced, trounces, cane, caning, caned, canes, thrash, thrashing, thrashed, thrashes, clobber, clobbering, clobbered, clobbers, spank, spanking, spanked, spanks, paddle, paddling, paddled, paddles, hit, hitting, hits, whack, whacking, whacked, whacks, pummel, pummelling, pummeling, pummelled, pummeled, pummeled, pummels, club, clubbing, clubbed, clubs, shoot, shooting, shot, shoots, suffocate, suffocating, suffocated, suffocates, beat, beating, beaten, beats.

The keyword-matching method uses the following rules to classify a sentence or document:

- KILL: If a *police* keyword *AND* a *kill* keyword appear in the same piece of text, classify it as a positive.
- ARREST: If a *police* keyword *AND* an *arrest* keyword appear in the same piece of text, classify it as a positive.
- FAIL TO ACT: If a *police* keyword *AND* an *intervention* keyword appear in the same piece of text, classify it as a positive. (This is a very simple rule-based method and we leave to future work to develop a keyword-based method that more adequately captures the *not* semantics of "did not intervene.")
- FORCE: If a *police* keyword *AND* an *force* keyword appear in the same piece of text, classify it as a positive.
- ANY ACTION: If a *police* keyword appears in a piece of text, classify it as a positive.

C.3 RoBERTa+MNLI.

We use the pretrained model from https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.md#pre-trained-models

C.4 ELECTRA+MS MARCO.

We use the pretrained model from the `sentence-transformers` package https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/information-retrieval#pre-trained-cross-encoders-re-ranker.

For IR, there are two standard architectures for scoring passages and queries: *cross-encoders* in which the architecture performs full attention over the pair and *bi-encoders* in which the passage and query are each mapped independently into a dense vector space (Luan et al., 2020). We chose a model with a cross-encoder architecture since these have been shown to consistently have higher performance (Thakur et al., 2021).

D Results

This section provides additional results beyond those included in the main paper, including results for document-level models, variants on the BM25 model, mean absolute error results to complement the Spearman correlations presented in the main paper, and the temporally aggregated results for all of the semantic event classes.

D.1 Document Level F1

To complement the sentence-level F1 metrics in the main paper, we present the document-level metrics in Table A4.

D.2 BM25 and Variants

In addition to the standard BM25 model reported in the paper, we tested several variants, including automatic term expansion using RM3 and manual term expansion using the same keywords from our keyword method. The results are shown in Table A5.

D.3 Spearman and MAE results

In the main paper, we report Spearman correlations between the daily count of gold standard events identified by our annotators. In Table A6 we also report the mean absolute error in daily event counts between our two models for each event class. We prefer Spearman correlations over MAE because the correlation is normalized between -1 and 1, while MAE tends to be higher for high-prevelance event classes.

D.4 Temporal Aggregates for All Event Classes

We report the temporal aggregate comparisons for all event classes in Figure A5 to supplement the figure in the main text showing results for the FAIL TO ACT class.

E Prototype span-based annotation schema

Before arriving at the annotations via natural language described in Section 4.2, we first attempted to gather *span-based* text annotations in order to collect more fine-grained details about police activity. In these prototype rounds, we first asked annotators to highlight spans in the text that answered "What action did police do?" Then given the action text-span they highlighted, we asked them to highlight spans for the following questions: "Police did the action *using what?"* "Police did the action *towards whom?*" "Where did the action occur?" "When did the action occur?" "Why did the action occur?"

There were several major barriers to this annotation schema that caused us to abandon the

	Document Level F1			
	Prop. Pos.	Keyword-Sent	Keyword-Doc	RoBERTa+MNLI
ANY ACTION	0.36	0.82	0.82	0.63
ARREST	0.10	0.68	0.50	0.63
FORCE	0.07	0.70	0.48	0.59
KILL	0.04	0.61	0.40	0.77
FAIL TO ACT	0.09	0.13	0.24	0.60

Table A4: Document level metrics for keyword matching and RoBERTa+MNLI model. "Keyword-sent" aggregates results from the sentence keyword matcher to the document level using an at least one threshold.

	Ranking Metrics: Ave. Precision (↑), PropRead@Recall95 (↓)			
	BM25	BM25+expn	BM25+RM3	BM25+RM3+expn
ANY ACTION	0.53, 0.85	0.76, 0.85	0.50, 0.83	0.65, 0.81
ARREST	0.68, 0.36	0.63, 0.21	0.58, 0.44	0.45, 0.43
FORCE	0.24, 0.43	0.52, 0.44	0.26, 0.42	0.67, 0.50
KILL	0.30, 0.29	0.22, 0.36	0.28, 0.33	0.36, 0.49
FAIL TO ACT	0.27, 0.77	0.24, 0.69	0.25, 0.72	0.21, 0.66

Table A5: Comparisons of BM25 and its variants. Here, *expn* means we use the manually-curated expanded keywords and append these to the original query as input into the model.

Event Class	Keyword (MAE)	RoBERTa+MNLI (MAE)
KILL	2.16	0.867
ARREST	2.10	4.45
FAIL TO ACT	3.25	1.48
FORCE	7.22	3.23
ANY ACTION	37.87	15.42

Table A6: Mean absolute error (MAE) between the human gold-standard annotations and the keyword-matching and RoBERTa+MNLI models.

span-based annotation approach for our current approach—pre-selecting semantic event classes of interest and having annotators give sentence classification labels. First, we were unable to resolve discrepancies in how much the annotators should highlight for given spans. Following the "argument reduction criterion" of Stanovsky et al. (2016), we asked annotators to "highlight as much as you need to answer the question but not more. If you can exclude a word from the highlighting without changing the answer to the question, you should exclude it." For example, in the text "Police suddenly attacked protesters with sticks" we expected annotators to highlight "suddenly attacked" versus just "attacked" because the former is a slightly different action. However, this criterion did not succeed in improving annotator agreement on span extents.

Furthermore, with span-based annotations, it was difficult to decide how to properly aggregate police actions (e.g. how do we automatically separate *suddenly attacked* from *attacked* from *did not attack*?) Had we been committed to span-based annotations, we may have had to develop much longer, more detailed guidelines, such as those from the Richer Event Description project (O'Gorman et al., 2016). We believe this approach—which requires more work in developing guidelines and training annotators in them—is less easily extensible to new problems and social science domains. Finally, in a training round, the action text spans that annotators did select were not very substantively interesting and worth the additional cost and effort on the part of annotators.[16]

[16] Substantively less-interesting police actions identified include *made, identify, placed, recorded, said, spotted, suggested, used.*

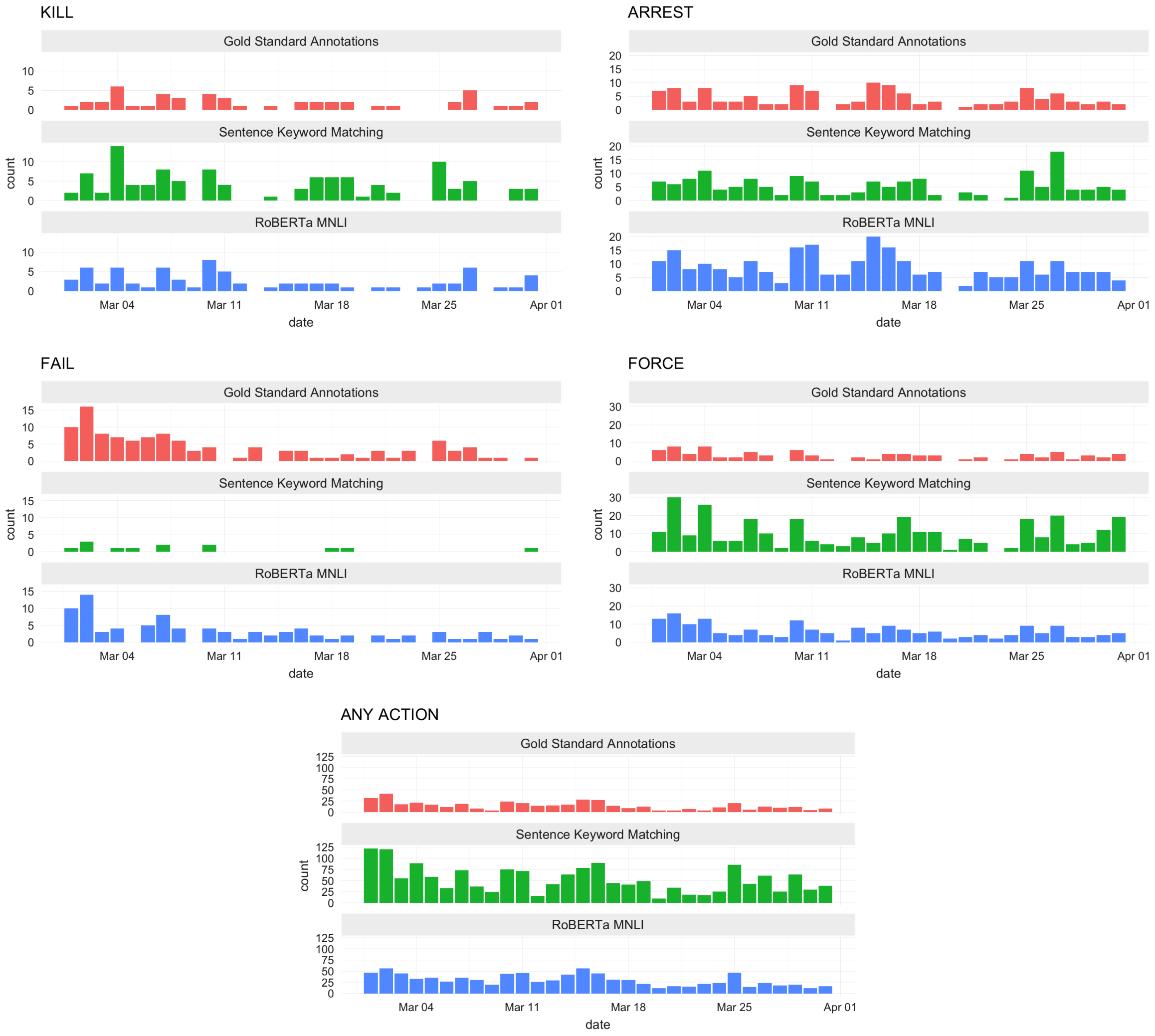


Figure A5: Temporal aggregate figures for all event classes comparing daily document counts of gold standard annotations, sentence keyword matching, and RoBERTa+MNLI.