DECEMBER 2024

ATAC
ASYMMETRIC THREATS ANALYSIS CENTER

The Global Terrorism Database

Experiments in Machine-Assisted Data Collection

Carlos R. Colon

Benjamin Evans

Margaret A. Hayden

START
NATIONAL CONSORTIUM FOR THE
STUDY OF TERRORISM AND RESPONSES TO TERRORISM

APPLIED RESEARCH LABORATORY FOR
INTELLIGENCE
AND SECURITY

UNIVERSITY OF
MARYLAND

ABOUT THE AUTHORS

The authors of this report are Carlos R. Colon (Faculty Specialist), Benjamin Evans (Senior Faculty Specialist), and Margaret A. Hayden (Faculty Specialist) of the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland. Dr. Jordan Boyd-Graber, Feng Gu, Zongxia Li, and Ishani Mondal of the University of Maryland's Computer Science (COMSCI) department; and Dr. Sean Doody and Dr. Michael Jensen of START provided valuable assistance in the development of this report.

Questions about this report should be directed to Carlos R. Colon at raffy@umd.edu.

ABOUT THE PROJECT

This interim report is part of the START project, "Global Terrorism Database Technical Updates," part of the Asymmetric Threat Analysis Center (ATAC), a joint program between START and UMD's Applied Research Lab for Intelligence and Security (ARLIS). ATAC is funded by the Department of Defense under award no. HQ003421F0481. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the Department of Defense.

ABOUT START

The National Consortium for the Study of Terrorism and Responses to Terrorism (START) is a university-based research, education and training center comprised of an international network of scholars committed to the scientific study of terrorism, responses to terrorism and related phenomena. Led by the University of Maryland, START is a Department of Homeland Security Emeritus Center of Excellence that is supported by multiple federal agencies and departments. START uses state-of-the-art theories, methods, and data from the social and behavioral sciences to improve understanding of the origins, dynamics, and effects of terrorism; the effectiveness and impacts of counterterrorism and CVE; and other matters of global and national security. For more information, visit www.start.umd.edu or contact START at infostart@umd.edu.

ABOUT ARLIS

The Applied Research Laboratory for Intelligence and Security (ARLIS), based at the University of Maryland College Park, was established in 2018 under the sponsorship of the Office of the Under Secretary of Defense for Intelligence and Security (OUSD(I&S)). As a University-Affiliated Research Center (UARC), ARLIS' purpose is to be a long-term strategic asset for research and development in artificial intelligence, information engineering, and human systems. ARLIS builds robust analysis and trusted tools in the "human domain" through its dedicated multidisciplinary and interdisciplinary teams, grounded both in the technical state of the art and a direct understanding of the complex challenges faced by the defense security and intelligence enterprise. For more information, visit www.arlis.umd.edu/about-arlis or contact ARLIS at info@arlis.umd.edu.

CONTENTS

Introduction

The Global Terrorism Database (GTD) research team, in collaboration with researchers from the University of Maryland's Computer Science (COMSCI) department, undertook a pilot project to evaluate the potential efficiency gains that could be achieved by relying more heavily on the use of artificial intelligence (AI) to compile the database. The primary motivation for this initiative was to reduce the time lag between the occurrence of real-time terrorist events and GTD data collection, as well as to reduce the costs associated with producing the data. A key area where it was hypothesized that the use of AI could have a positive impact is the time required for human analysts to identify and code events that meet the GTD's inclusion criteria.

To evaluate the potential of using AI more extensively in the collection of the GTD, we conducted two experiments using different natural language processing (NLP) methodologies (i.e., automated and computational techniques for extracting information from text). The first experiment focused on automatically identifying individual terrorist attacks and clustering together all documents referring to the same events from a pool of potentially relevant news articles. The second experiment aimed to automatically extract detailed information about each attack. Our core research question was whether AI tools could correctly identify unique terrorism events from global news sources and then correctly map the relevant features of the events to the variables included in the GTD.

While the automated NLP techniques we tested have not yet reached human-level accuracy, our findings indicate that they can reduce the manual workload and the time required for human annotation. By adopting embedding-based methods (described below) for document clustering and integrating language models (LMs) into the data coding pipeline, the use of additional AI tools can achieve a more efficient workflow. However, our results suggest that a fully automated GTD would not achieve an acceptable level of accuracy, indicating the continued need for a human-AI hybrid collection methodology.

This report is organized as follows: The next section provides background on the GTD workflow, detailing our current automation pipeline and the human effort involved in identifying and coding events. Then, we describe our experiments and discuss the results. Finally, we outline our conclusions and the next steps for future research.

Background

The GTD contains over 200,000 records of terrorist events that have occurred around the world since 1970. To maintain the database, the team adopts a hybrid strategy in which automated techniques are used to narrow a pool of source documents, which are then passed to human subject matter experts for event identification, detailed coding, and quality control. Figure 1 illustrates the GTD's data collection process.

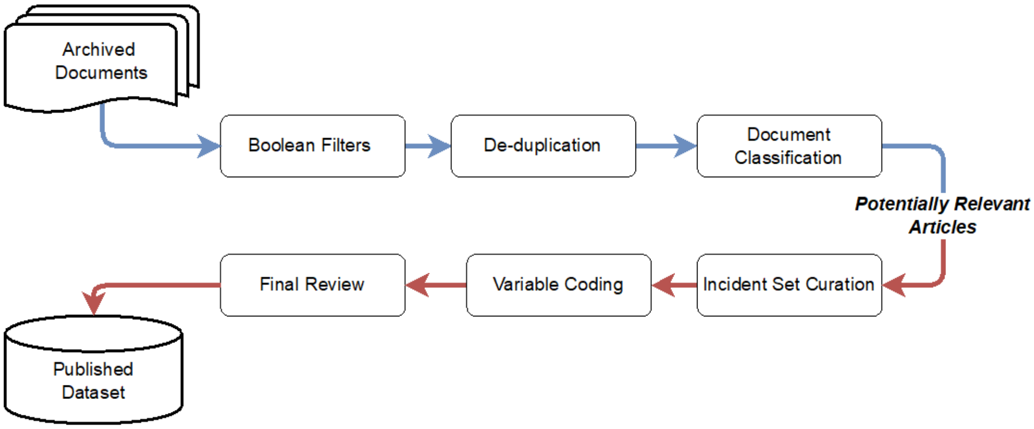

Figure 1: The GTD data collection workflow.

Automated Source Collection

Currently, the GTD downloads an average of 2.2 million news articles per day from commercial platforms, such as LexisNexis and BBC Monitoring.[1] The purpose of the automated source collection stage is to identify potentially relevant news articles, enabling the research team to focus on a smaller, manageable subset of articles related to political violence.

The source collection stage begins by filtering the collected news articles based on Boolean search terms, developed by subject matter experts, to retain documents most likely to include content about terrorism. For example, search strings combine terms like “arson” or “set fire” with “gunmen” or “militant,” while excluding contexts such as “film” or “television.” Even after this initial filtering, however, the remaining articles include a substantial number of duplicate news reports, as well as articles that are, at best, tangentially related to terrorist attacks.

To eliminate duplicate articles, the team uses an NLP technique called term frequency-inverse document frequency (TF-IDF). TF-IDF creates a vector of term weights for each article, where the terms most important to an article have the highest weights. The TF-IDF vectors can then be used to identify articles with very similar textual representations using a metric called cosine similarity.[2] Articles exceeding a predetermined cosine similarity threshold are flagged,[3] and those from less credible sources are removed from the pool in favor of those from more trusted outlets.[4]

[1] For example, between October 2020 and December 2021, the GTD archived and pre-processed 949,239,073 source documents—averaging 63.3 million per month or 2.2 million per day. See Miller and Wingenroth (2022).

[2] Cosine similarity ranges from -1 to 1, with values closer to 1 implying more similarity. In our case, articles with similar TF-IDF vectors, and therefore similarly weighted terms, will have higher cosine similarities.

[3] In our case, articles with a cosine similarity score of 0.825 or higher are flagged as duplicative.

[4] The research team assigns source validity scores to media outlets. Content sourced from social media platforms like Twitter or Facebook are assigned a validity score of 1, while articles from trusted outlets like *Reuters* or *The Washington Post* are assigned a score of 3. Articles with higher source validity scores are given priority over those with lower scores.

After deduplication, the pool of articles averages about 80,000 per month. For the final step of the automation stage, we use machine learning tools for further refinement. We begin by standardizing the data, which involves cleaning raw text, removing stopwords (e.g., frequently occurring words like "a," "an," "its" and so on that do not add analytic value), and stemming words (i.e., reducing words to root form to simplify the corpus vocabulary, such as "fighting" to "fight"). After preprocessing the text, documents are then classified into two distinct categories: those that are likely related to terrorism incidents and those that are not. For this task, we employ a Support Vector Machine (SVM) model that takes a TF-IDF matrix of articles as an input, assigns relevancy scores to each article, and flags highly relevant documents for manual annotation.[5] After completing this classification step, the number of articles selected for manual review is reduced to around 10,000 to 13,000 per month.

Human Annotation

The human annotation stage involves three key steps: (1) creating incidents; (2) coding variables; and (3) conducting a final quality control review. In the first step, researchers identify terrorist attacks and group together all articles that describe the same events. To do this, subject matter experts read thousands of news articles that have been classified as relevant by the SVM model to systematically identify unique terrorism events that satisfy the GTD inclusion criteria. They then record preliminary information about each event in a structured data table and attach the supporting source documents. This allows other researchers to then read the documents and code over 100 variables for each event that are derived from the GTD codebook.

Articles are processed on a per-month basis. After incidents have been created from a monthly batch of global news articles, detailed incident coding begins. In this stage, researchers are organized into small, specialized teams. Each team focuses on different coding domains, such as: location; perpetrators; targets; weapons/tactics; and casualties/consequences. These teams read the attached source documents to extract incident information pertaining to GTD variables. Finally, the coded records undergo a final quality review. In this phase, researchers review the data for consistency, add additional summary variables, and remove any records that fail to meet the inclusion criteria. Once quality control is complete, the data is published online.

Ongoing Challenges in the GTD's Data Collection Process

Despite the automated filtering steps in the GTD's data collection pipeline, a significant portion of the articles that are passed on to researchers for manual annotation remains irrelevant or redundant. For instance, during the review of April 2021 articles, the GTD team read 13,747 news articles, of which only 16.4% (2,255) contributed to the creation of 1,027 terrorist attacks.[6] Similarly, in December 2020, out of 10,711 reviewed articles, approximately 10.6% (1,139) were linked to incident records.

[5] Relevancy scores are assigned on a scale of -2 to 2, with 2 corresponding to an article that almost certainly contains information about an incident of terrorism. A score of-2 corresponds to a story that satisfies the Boolean filters but does not necessarily describe an act of terrorism.

[6] Initially 1,084 records were created, but 57 of them were deleted for various reasons. By the end of the GTD triaging process, only 794 incidents were included in the final data for April 2021 after 233 incidents were withheld due to insufficient sourcing. See Miller and Wingenroth (2022, p. 10-14).

The review of irrelevant and redundant content by human analysts causes delays in the publication of new data and is a significant contributor to the high costs associated with compiling the dataset. In 2016, as available funding for terrorism event data collection was beginning to decline, we identified the increased use of AI in the GTD's data collection workflow as a way to potentially address these concerns. We conducted a pilot study comparing the incident identification accuracy of fully automated, fully human-coded, and human-machine hybrid collection techniques used by political violence researchers (Wingenroth et al. 2016). We found that the technology used in fully automated terrorism datasets at the time was not capable of producing data accurate enough to be used by the scientific community. In comparison to human-coded and human-machine collected data, the fully automated data we reviewed were plagued by incredibly high rates of false positive, false negative, and duplicate cases. For example, one of the fully automated datasets we reviewed at the time included more than 900 successful assassinations of former President Obama.

Previous work conducted by other researchers found similar results. One study showed that as event attributes grow more complex and detailed, machine coding accuracy declines (King and Lowe 2003).[7] Another study demonstrated significant discrepancies between machine-coded and human-coded event data, largely stemming from geolocation errors (Hammond and Weidmann 2014). Other researchers found that fully automated event datasets were prone to data inflation due to duplicates, misclassification of irrelevant events, and inclusion of events that never actually occurred (Miller et al. 2022; Raleigh, Kishi, and Linke 2023).

While the results of our original pilot study convinced us that GTD's human-machine hybrid approach was likely achieving all the efficiency gains that automation could provide at the time, AI technology has vastly improved in the years since we completed the project, especially when it comes to the rapid development of large language models (LLMs) like ChatGPT. Given these improvements, we decided to revisit the question of whether the increased use of AI tools could improve the efficiency and cost-effectiveness of the GTD. To answer these questions, we conducted two experiments. The first experiment directly addressed the challenge of incident identification by testing automated methods to identify individual terrorist attacks from news articles and cluster together all reports about the same events. By doing so, we aimed to reduce the burden on researchers and increase the efficiency of incident set creation.

As we noted above, after relevant events are identified for inclusion in the GTD, researchers must reread the source documents used to create the events in order to code detailed information about the attacks, including where they occurred, what weapons were used, who perpetrated the attacks, and how many people were killed or injured. Therefore, the second experiment focused on accelerating this phase of the data collection process by leveraging language models to automatically extract key GTD variables from the text of news articles.

Experiment 1: Incident Set Curation

[7] For example, King and Lowe (2003) showed that while machine coding was comparable in accuracy to human coders overall, its accuracy diminished when classifying more nuanced event types—dropping to 25–55 percent for detailed categories like political graffiti—though it achieved better results of around 55–70 percent for more general event types such as protest demonstrations (see Table 2, p. 631).

Technical Design

The objective of the incident set curation experiment was to examine whether improvements can be made to the automated processes currently used by the GTD, as well as to determine if a fully automated approach can replicate human efforts in generating incident sets. To this end, the COMSCI team was provided with 19,295 triaged news articles (i.e., articles that have been read and reviewed by human annotators) containing incident metadata from October 2020 through December 2020. Using these data, the COMSCI team trained an LM classifier called RoBERTa (Liu et al. 2019) to predict whether or not an article is relevant in an effort to enhance the performance of the GTD's classification methodology.[8] These documents were split into training, testing, and validation sets at an 80/10/10 ratio (i.e., 80% of the articles were used for training, 10% were used for testing, and 10% were used for validation).

Next, the trained RoBERTa classifier was tasked with making predictions about 7,941 unseen documents from a "holdout" dataset of February 2022 news articles. GTD researchers were randomly assigned 500 documents predicted to be relevant to create a reference set of incidents. From these 500 articles, the GTD research team generated 371 incidents, with an average of 1.8 articles per incident. This reference set of incidents served as the "gold standard" for evaluating the subsequent performance of the automated methods used for incident identification.

The next step in the experiment used automated methods to group articles that pertain to the same terrorist event into clusters to compare how they perform relative to human incident curation. Four different clustering methodologies were tested, each with the goal of maximizing the similarity of the articles within their group:

- **Baseline TF-IDF:** The GTD does not currently attempt to cluster articles into incident sets. However, the GTD currently uses the Nearest Neighbors algorithm (with a radius of 1.35) on top of a TF-IDF matrix to organize articles into relevant groupings for manual incident identification. Therefore, this serves as the baseline comparison for the other methods.
- **Embedding:** The COMSCI team used state-of-the-art BAAI General Embeddings (Xiao et al. 2023) to transform the articles into "document embeddings"—outputs of a language model that represent semantic and textual features of texts quantitatively and accurately. Each article receives an "embedding" and can be compared to other article embeddings for semantic similarity. The COMSCI team trained the BAAI model to determine the optimal level of similarity for assigning a given pair of articles to the same cluster. If a pair of articles meet the optimal similarity threshold learned during training, then they are assigned to the same cluster.
- **Language Model Classification (LM-CLS):** OpenAI's GPT-4o-mini LLM (OpenAI 2023) was deployed to determine if two documents described the same incident. Specifically, the COMSCI team prompted the model to give a binary answer: "yes" if two article texts described the same event, and "no" otherwise.
- **Language Model Segmentation (LM-SEG):** A recurring challenge in incident curation is the issue of articles mentioning multiple incidents. In this final clustering task, the COMSCI team prompted GPT-4o-mini to segment documents—i.e., break them into chunks according to

[8] The RoBERTa model was trained on articles that were already classified as relevant by the GTD's existing SVM model to determine if the use of a neural language model can refine existing GTD methodologies.

discrete events—prior to being classified by LM-CLS. In this procedure, if a document described multiple incidents, the model separated incident-specific texts into their own segments. Therefore, LM-SEG should increase the accuracy of article classification assuming that GPT-4o-mini segments documents effectively.

The COMSCI team evaluated the performance of the TF-IDF, Embedding, LM-CLS, and LM-SEG procedures by comparing their generated incident sets to the gold standard reference set manually compiled by GTD researchers using precision, recall, and F_1 scores. In this context, precision is the fraction of correctly predicted positive cases over all of the positive predictions made by the model, and recall is the fraction of correctly predicted positive cases made by the model over the number of true positives in the dataset.[9] The F_1 score combines insights from both the precision and recall scores to help us see which method performed the best overall, with higher scores indicating better performance.

These methods must make trade-offs between Type I and Type II errors. In this context, a Type I error (false positive) refers to instances where the automated method incorrectly identifies a source document as relevant to an incident set when it is not. Conversely, a Type II error (false negative) happens when a relevant document fails to be identified as part of an incident set. In terms of impact to workflow, Type I errors lead to an increased workload for researchers, who have to sift through irrelevant news articles during incident coding. Type II errors, on the other hand, risk researchers missing relevant articles, which threatens the accuracy of the data. Without these relevant articles, researchers may potentially overlook important information or, worse, fail to capture entire incidents without a systematic way of knowing what is missing from the dataset. This experiment yielded three primary findings:

Results

	Precision	**Recall**	**F_1**
TF-IDF	0.19	0.09	0.10
Embedding	**0.89**	0.51	0.59
LM-CLS	0.36	0.35	0.35
LM-SEG	0.65	**0.66**	**0.63**

Table 1: Embedding has the highest precision of 0.89. However, LM-SEG has the highest recall and overall F_1 score.

Automated clustering cannot yet match human performance. The LM-SEG method performed the best overall, obtaining the highest F_1 score of 0.63 (Table 1). However, despite being significantly better than the TF-IDF baseline, LM-SEG still includes over 30 percent of Type I and Type II errors, therefore failing to match the performance of trained human coders. It is worth noting that all methods generated fewer incident sets than the human-identified reference set. This undercounting may be in part due to the conditions of the experiment, in which an individual article—or segment, in the case of LM-SEG—could only belong to one incident set. Further testing

[9] Here, a “positive” prediction means an article was predicted to belong to a given incident set. If for some cluster *X* the set of true articles is [1, 3, 5, 7] and the model predicts [1, 3, 4, 6], then the precision is equal to 50% because half of the model’s positive predictions were correct. Likewise, for recall, the model only successfully predicted two positive cases out of four—[1, 3]—and also achieved a score of 50%.

should examine loosening this restriction to allow articles to be part of multiple incident sets while ensuring that each incident set refers to a distinct event.

Automatically generating incident sets with human-level accuracy remains a challenge. Though automated incident creation is not yet feasible, Embedding and LM-SEG are still valuable options to replace TF-IDF in the current GTD workflow. These methods would likely provide improved performance in suggesting relevant documents and presenting more cohesive sets of articles for researchers during manual incident identification, but further qualitative analysis is needed to confirm this hypothesis.

Embedding effectively captures the semantic relationships between documents. Unlike TF-IDF, which provides limited insights for creating incident sets, Embedding significantly improves performance, achieving 89 percent precision. In our experiments, TF-IDF attached less than 10 percent of relevant documents, and less than 20 percent of attached documents are relevant. In contrast, Embedding reduces false negatives by 82 percent, identifying roughly half of the relevant documents. Overall, Embedding achieves an F_1 score of 0.59 (Table 1).

Segmentation improves clustering accuracy. While LM-CLS also achieves higher precision, recall, and F_1 than TF-IDF, the challenge of organizing articles containing multiple incidents hindered its performance. With a precision of 0.36 and a recall of 0.35, LM-CLS falls short of Embedding by a large margin (Table 1). However, by breaking up documents into discrete segments before clustering, LM-SEG achieves over 60 percent in both precision and recall, outperforming LM-CLS and all other methods. Moreover, many sets curated by LM-SEG overlap with the incidents in the gold standard reference set. Of the 371 incident sets identified manually by the GTD research team, 203 were produced identically by LM-SEG. In contrast, LM-CLS yields an overlap of 105 incident sets, Embedding yields 66, and TF-IDF yields only 12 matching sets. Given the prevalence of source articles that discuss multiple incidents—including across different countries—segmentation appears to be an important step for clustering articles most effectively.

Experiment 2: Variable Coding

Technical Design

To study whether language models can alleviate human effort in variable coding, we used GPT-4o mini to extract GTD variables from the news articles. We separated the incident sets into two groups:

- **Manual:** The researchers coded the variables without assistance, serving as the control group.
- **Optional:** The researchers saw the extracted variables from the attached documents for each variable and could populate the fields with the suggested values.

To obtain more robust results, we duplicated a subset of incidents and assigned them to three groups of human annotators. Each group coded 212 incident sets, evenly divided between the manual and optional coding settings, resulting in a total of 636 coded incidents—318 under each setting. For each incident, human annotators coded nine variables. No team coded the same incident under both settings. To understand how misinformed incident sets—i.e., incident sets containing irrelevant articles—affected human annotation, we replaced some incidents from the reference set with their nearest neighbors from LM-SEG. Overall, our 636 coded incidents consisted of 371 unique incident sets: 93 from the reference set, 75 misinformed sets from LM-SEG, and 203 overlapping sets between both LM-SEG and the gold standard reference set.

We analyzed the time taken for each researcher to complete the coding and evaluated the agreement between the extracted variables and the coded variables by human annotators. Because string match methods suffer from Type II errors (e.g., spelling and formatting), we use additional equivalence metrics. Specifically, we use the following methods to determine the agreement:

- **Normalized Match (NM):** The == ("exactly equal to") operator checks if two variables are the same after string normalization by removing delimiters and converting variables to lower case.
- **BERT:** A RoBERTa model fine-tuned on human-labeled and synthetically generated string matching datasets to measure the similarity between two string embeddings.
- **PEDANTS:** A model that uses an optimized learned F_1 and TF-IDF encoding to measure agreement between two strings (Li et al. 2024).

Results

	Human-Only	LM-Only	Overlap	Average
Manual	271 (160)	233 (227)	220 (145)	236 (169)
Optional	260 (185)	175 (202)	173 (132)	197 (169)
Average	265 (174)	203 (216)	197 (140)	217 (170)

Table 2: The average time, in seconds, taken by human annotators to complete annotations for different types of incident sets (along with standard deviations in parentheses). The table columns are incident sets, where Human-Only refers to the 93 incidents selected from the reference set, LM-Only refers to the 75 misinformed incident sets from LM-SEG, and Overlap refers to the 203 equivalent incident sets produced by both human annotators and LM-SEG.

This experiment produced several important results:

Optional answers reduce annotation time. Researchers consistently spend less time coding when they have access to variables extracted by GPT-4o-mini. Table 2 demonstrates that the optional setting, on average, decreases annotation time by approximately 16 percent compared to the manual setting. While the time saved is marginal in absolute terms, it allays any potential concern that introducing extracted variables, in addition to the article texts, might increase the workload for researchers. Moreover, this finding is consistent across all types of incident sets. When annotators and LM-SEG concur on incident sets, annotators spend the least time coding. Even when misinformation is present (i.e., incorrect articles are attached) in the incident sets generated by LM-SEG, extracted variables still reduce annotation time by 25 percent.

Pre-populated variables provide high utility. Overall, annotators used extracted variables 66 percent of the time (Table 3). The most frequently selected variable, Country, had a likelihood of selection exceeding 92 percent. Even the least frequently selected variable, Location, was chosen over 40 percent of the time, proving somewhat helpful for coding. In some instances, extracted variables were not chosen by the researchers because they were incomplete or imprecise. For example, the extraction may have included some, but not all, of the weapons used in an attack, or it may have identified the first-level administrative division but failed to capture the specific town in which an event occurred. This suggests that fine-tuning the extraction method may yield a higher selection frequency.

Variable	Selection Frequency (%)	N
Country	92.1	278
Location	40.1	278
Target	64.5	265
Perpetrator	78.5	209
Generic Attack	58.5	241
Generic Weapon	73.7	218
Specific Weapon	61.3	194
Kills	79.4	106
Wounds	67.8	97
Overall	66.0	1886

Table 3: Total number of non-NA (i.e., not empty) variables and their selection frequency in the Optional setting. The N column corresponds to the number of times that the corresponding variable had a value (i.e., the answer was not NA). The Selection Frequency (%) column denotes the rate at which the extracted variable was chosen by the human coder.

Additionally, in all three incident set types, the optional setting showed higher agreement than the manual setting, suggesting that annotators effectively utilized the extracted variables (Figure 2). Standardizing the extracted variables, especially according to the existing structure of the GTD data, would likely further increase intercoder agreement and reduce the presence of duplicate outputs that contain the same content but differ in formatting.

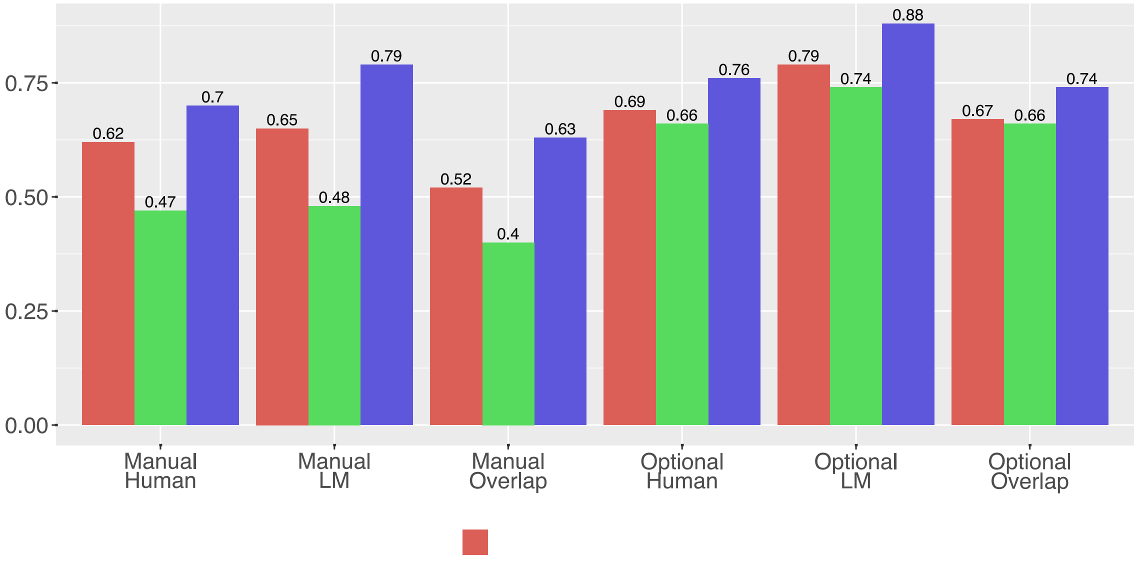

Figure 2: Agreement grouped by incident set type and setting. Human annotators agree more with the LM-extracted variables under the optional setting, in which extracted variables are present. This suggests that the extracted variables are helpful for coding. Additionally, the extracted variables are more useful in LM-generated incident sets, where conflicting information is more prevalent.

Conclusions and Future Research

While the automated approaches explored in this pilot project have not yet achieved human-level accuracy, our findings highlight several promising avenues for improving the GTD data collection methodology.

Replace TF-IDF with Embedding

Statistical models like TF-IDF struggle to capture the semantic relationships between documents (Bengio et al. 2003; Mikolov et al. 2013), risking the exclusion of documents that are textually different but semantically similar. In contrast, the Embedding and LM-SEG methods both significantly outperformed TF-IDF in generating incident sets. Although LM-SEG achieved slightly higher overall performance than Embedding, its computational cost grows significantly with the number of documents, making it less scalable. Therefore, the most practical improvement to the current GTD pipeline is to replace TF-IDF with Embedding when clustering articles for researchers prior to manual incident identification, balancing improved accuracy with computational feasibility.

The improved performance of the LM-SEG method compared to LM-CLS affirms the utility of segmentation for accurately clustering source articles. However, using a commercial language model like OpenAI's GPT-4o-mini for segmentation is currently cost prohibitive, and we were unable to identify an open-source model that performed segmentation with sufficient accuracy for the purposes of this pilot study. Nevertheless, successful text segmentation is a necessary development in order to pursue the ultimate goal of automatic incident identification. Accurate segmentation could also improve variable extraction by allowing the model to focus solely on the most relevant section of text within an article.

Enhance Relevance Classification with RoBERTa

Although the primary focus of our experiments was generating and coding incident sets, our tests also revealed that a RoBERTa-based relevance classifier can achieve near-expert-level performance. Of the random sample of 500 articles predicted to be relevant by RoBERTa, researchers used 358 (71.6%) to create incidents during the first experiment, suggesting that the classifier performed well in identifying relevant documents. With an accuracy of 95 percent—representing a substantial 76 percent improvement over the current classifier's accuracy of 54 percent—RoBERTa stands as a valuable replacement, or supplement, to the existing SVM model. Incorporating RoBERTa into the GTD data collection pipeline as a direct replacement or a secondary filter should reduce the volume of irrelevant articles passed on to researchers for human annotation.

Refine Language Model-Assisted Variable Coding with Confidence-Based Answers

While language model assistance only slightly reduced coding time in our experiment, it holds promise for enhancing variable coding. Further refinement in LM-driven information extraction could substantially increase its utility. Current challenges include issues with specificity (e.g., extracting a broader administrative region instead of a village), extracting the incorrect answer despite identifying the correct source text, and confidently extracting conflicting or ambiguous information.

A logical avenue for enhancing variable extraction is the use of confidence-based answers. Implementing a confidence-based answering system could take a variety of forms. At a minimum, the extraction model could abstain from suggesting low-confidence answers (Feng et al. 2024) and/or display a confidence score alongside the extracted variable, reducing noise while providing researchers with more transparent guidance to make reliable and accurate coding decisions. With enough success, a more comprehensive approach could automatically code variables above a high confidence threshold, bypassing human coding for straightforward events. Alternatively, LM-powered variable coding could produce lower-quality but more immediate data for users, delivering rapid reports containing a limited number of top-line variables until the fully coded gold standard data is available.

Guide Extractions with Incident-Level Context and Retrieval-Augmented Generation

Another potential strategy for enhancing variable coding is transitioning from document-level to incident-level extraction. For our experiment, each article's information was processed independently, which led to inaccuracies when a single article referenced multiple incidents or when multiple articles described the same incident from different perspectives. Instead, cross-referencing and synthesizing information from all articles attached to each incident could produce more accurate and contextually informed extractions. This broader, incident-level perspective could help resolve conflicting information, better combine details from related sources, and ultimately improve the precision of the extracted variables.

The integration of retrieval-augmented generation (RAG) also presents an opportunity for development. Rather than relying solely on the model's internal knowledge, RAG incorporates authoritative external references—such as previously annotated incidents or variable definitions and coding nuances from the GTD codebook—into the extraction process. For example, before determining the number of casualties in an attack, the model could consult this external knowledge base maintained by GTD researchers to learn how to handle conflicting source reports. Therefore, RAG ensures that the model's outputs are guided by up-to-date, context-specific data, resulting in more accurate and reliable variable coding.

Based on the results of our experiments, implementing these adjustments to the clustering and classification models within the automated source collection stage can move the GTD's data collection process closer to a more efficient and scalable pipeline. In addition, the potential enhancements identified in this report, particularly the development of text segmentation and a confidence-based answering system for variable coding, could further upgrade the GTD's data collection methodology, combining technological capabilities with human subject matter expertise to achieve more rapid data collection without sacrificing accuracy.

References

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. "A Neural Probabilistic Language Model." *Journal of Machine Learning Research* 3: 1137–1155. https://dl.acm.org/doi/10.5555/944919.944966.

Feng, Shangbin, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. "Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics.* Bangkok: Association for Computational Linguistics. 14664–14690. https://doi.org/10.18653/v1/2024.acl-long.786.

Hammond, Jesse, and Nils B Weidmann. 2014. "Using machine-coded event data for the micro-level study of political violence." *Research and Politics* 1(2): 1-8. https://doi.org/10.1177/2053168014539924.

King, Gary, and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57(3): 617-642. https://doi.org/10.1017/S0020818303573064.

Li, Zongxia, Ishani Mondal, Yijun Liang, Huy Nghiem, and Jordan Lee Boyd-Graber. 2024. "PEDANTS: Cheap but Effective and Interpretable Answer Equivalence." *arXiv*. https://doi.org/10.48550/arXiv.2402.11161.

Liu, Yinhan, 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*. https://doi.org/10.48550/arXiv.1907.11692.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality." *arXiv*. https://doi.org/10.48550/arXiv.1310.4546.

Miller, Erin, and Brian Wingenroth. 2022. *Global Terrorism Database: Real-time Data Collection Pilot Evaluation.* College Park: START. https://www.start.umd.edu/publication/global-terrorism-database-real-time-data-collection-pilot-evaluation.

Miller, Erin, Roudabeh Kishi, Clionadh Raleigh, and Caitriona Dowd. 2022. "An agenda for addressing bias in conflict data." *Scientific Data* 9 (593): 1-6. https://doi.org/10.1038/s41597-022-01705-8.

OpenAI. 2023. "GPT-4 Technical Report." *arXiv.* https://doi.org/10.48550/arXiv.2303.08774

Raleigh, Clionadh, Roudabeh Kishi, and Andrew Linke. 2023. "Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices." *Humanities and Social Sciences Communications* 10 (74): 1-17. https://doi.org/10.1057/s41599-023-01559-4.

Wingenroth, Brian, Erin Miller, Michael Jensen, Omi Hodwitz, and Kieran Quinlan. 2016. "Event Data and the Construction of Reality." Paper presented at the SBP-BRiMS2016 Grand Challenge.

Xiao, Shitao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. "C-Pack: Packed Resources For General Chinese Embeddings." *arXiv*. https://doi.org/10.48550/arXiv.2309.07597.

ATAC

ASYMMETRIC THREATS ANALYSIS CENTER

National Consortium for the Study of Terrorism and Responses to Terrorism (START)
University of Maryland, College Park, MD 20740
infostart@umd.edu
www.start.umd.edu