

Near-real-time Earthquake-induced Fatality Estimation using Crowdsourced Data and Large-Language Models

Chenguang Wang ^{*}
 Stony Brook University
 Stony Brook, New York, USA

James Hou
 Stony Brook University
 Stony Brook, New York, USA

Davis Engler ^{*}
 U.S. Geological Survey
 Golden, Colorado, USA

David J. Wald
 U.S. Geological Survey
 Golden, Colorado, USA

Xuechun Li
 Johns Hopkins University
 Baltimore, Maryland, USA

Kishor Jaiswal
 U.S. Geological Survey
 Golden, Colorado, USA

Susu Xu
 Johns Hopkins University
 Baltimore, Maryland, USA

ABSTRACT

When a damaging earthquake occurs, immediate information about casualties (e.g., fatalities and injuries) is critical for time-sensitive decision-making by emergency response and aid agencies in the first hours and days. Systems such as Prompt Assessment of Global Earthquakes for Response (PAGER) by the U.S. Geological Survey (USGS) were developed to provide a forecast of such impacts within about 30 minutes of any significant earthquake globally. However, existing disaster-induced human loss estimation systems often rely on early casualty reports manually retrieved from global traditional media, which are labor-intensive, time-consuming, and have significant time latencies. Recent approaches use keyword matching and topic modeling to identify human casualty-relevant information from social media, but tend to be error-prone when dealing with complex semantics in multi-lingual text data, and parsing dynamically changing and conflicting human death and injury number shared by various unvetted sources in social media platforms.

In this work, we introduce an end-to-end framework to significantly improve the timeliness and accuracy of global earthquake-induced human loss forecasting using multi-lingual, crowdsourced social media. Our framework integrates (1) a hierarchical casualty extraction model built upon large language models, prompt design, and few-shot learning to retrieve quantitative human loss claims from social media, (2) a physical constraint-aware, dynamic-truth discovery model that discovers the truthful human loss from massive noisy and potentially conflicting human loss claims, and (3) a Bayesian updating loss projection model that dynamically updates the final loss estimation using discovered truths. We test the framework in real-time on a series of global earthquake events in 2021 and 2022 and show that our framework effectively automates the retrieval of casualty information faster but with comparable accuracy to those now retrieved manually by the USGS.

1 INTRODUCTION

Short-notice disastrous events, such as earthquakes, often cause considerable tremendous human costs, including fatalities, injuries, and displaced persons [19]. Immediate information concerning casualties after such natural disasters plays an important but challenging role in the post-disaster response. Traditional casualty reports

are often obtained from emergency response and rescue teams in the field, which often takes weeks to months [22, 33]. In the interim, many global emergency aid agencies and first responders currently refer to the casualty estimates provided by the Prompt Assessment of Global Earthquakes for Response (PAGER) system, developed by the U.S. Geological Survey (USGS) [18]. The PAGER reports provide a range of possible human fatalities (and economic impacts) within 30 minutes of any significant global earthquake, by updating an empirical fatality model using early casualty and injury reports and updated ground shaking maps [22]. However, in current practices, these early reports of human deaths and injuries are often manually retrieved from traditional media like Reuters or CNN, which is labour-intensive and has significant time latency.

Compared to traditional information sources with time delays, social media platforms provide access for the masses to directly share their feelings and observations concerning an evolving disaster, thereby providing potentially timely and useful onsite data compared to traditional media and field surveys [11, 14]. For example, we found that the first social media post reporting human deaths in 2022 M7.6 Papua New Guinea Earthquake is from a personal account within the macroseismic zone, posting a snapshot of a local community Facebook forum in Wau, Papua New Guinea, indicating 3 reported deaths. Existing social media scraping approaches mainly focus on categorizing the relevance level of text data instead of extracting exact casualty reports (e.g., death number, injuries number) and their locations. Researchers explored classic machine learning techniques, such as Support Vector Machines, Convolutional Neural Nets, and logistic regression, in combination with pre-trained disaster/social media post word embeddings to categorize relevant information [1, 16, 27, 32]. For example, CrisisNLP is a crisis informatics effort that leverages social media to collect disaster-related Twitter data and uses classifiers with traditional topic modeling [3]. However, these approaches mainly focus on categorizing the gathered articles or text data, instead of extracting exact numbers and locations. In addition, they are not robust against the highly complex and noisy social media text with large amounts of misinformation and ambiguities, due to the limited capability of traditional word embedding and topic modeling methods. Besides, previous work mainly focuses on English and Chinese data,

^{*}Equal Contributions

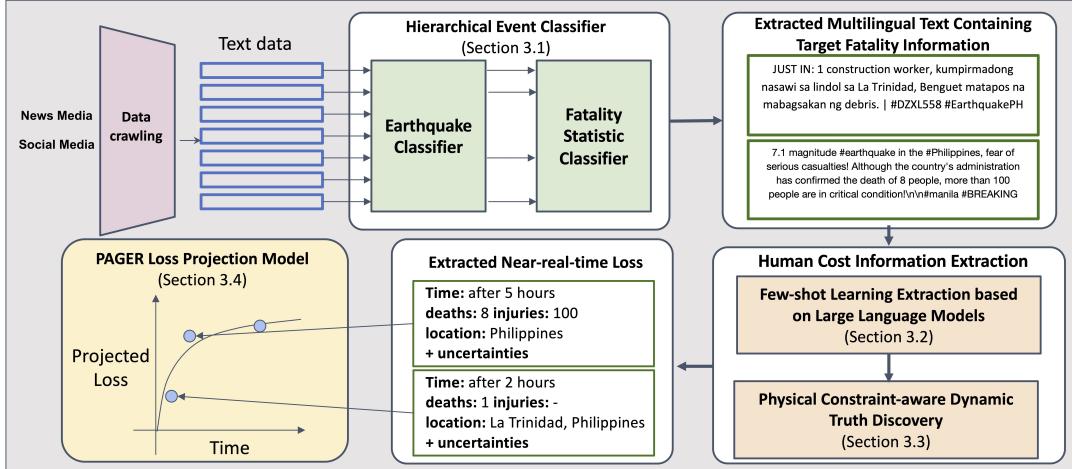


Figure 1: Overview of our framework’s design and application. Texts are crawled using keyword searches from social media and news sources. The data are first filtered with Hierarchical Event Classifier to extract texts that are highly possibly related to seismic human fatality (Section 3.1). Structured data (human fatality number and injury number) are further extracted from these filtered texts (Section 3.2). Furthermore, a physical constraint-aware dynamic truth discovery algorithm is introduced to cross-validate and discover the ground truth human fatality/injury number from various claims by various sources/accounts (Section 3.3). The output is eventually incorporated into an earthquake loss prediction model built on Bayesian updating to project the total human loss induced by an earthquake (Section 3.4).

neglecting the abundance of multilingual data present in global earthquakes.

To fill these gaps, our objective is to achieve automatic retrieval of the exact number of earthquake-induced human losses (death and injury) from multi-sourced social media and traditional media platforms for global events. We identify three important challenges posed by our objective. First, the multi-lingual text data shared by people around the world often contain a variety of complex semantics. For example, use of abbreviations and jargons [12, 15], or recollections of casualties from past events (e.g., recall of the 2010 Haiti when searching the 2021 Haiti Earthquake) and co-occurring unrelated non-earthquake emergency events (e.g. COVID). Second, the text data from different sources often contain incorrect and potentially conflicting information from a large number of unvetted sources. For example, a piece of misinformation on social media, saying 16 deaths in one hour after 2022 M7.6 Papua New Guinea Earthquake, has been widely circulated by many verified public media accounts but was later claimed to be misinformation. The unknown reliability of various data sources makes it challenging to extract accurate information. Third, the reported human costs dynamically evolve with heterogeneous region-specific patterns tied to resource availability, meaning the ground-truth value is changing as well. However, information spreading on social networks often takes time and exhibits delay patterns, thus, delayed data often appear more prominently than the latest, more accurate data. This largely constrains the timeliness of information retrieval and poses additional difficulties when cross-sourcing information for verification. Moreover, due to time sensitivity, it is impossible to have experts label large amounts of extracted text data for fine-tuning and adapting the information retrieval models to data reporting patterns specific to the earthquake of interest.

To address these challenges, we introduce a novel, near-real-time, end-to-end framework that can automatically retrieve accurate human casualty information from multiple data sources and adaptively by integrating Large Language Models (LLMs) and dynamic truth discovery, as shown in Figure 1. Specifically, this work makes the following contributions:

- (1) We develop a hierarchical event-specific disaster data extraction framework that leverages a multilingual event classifier, prior knowledge of LLMs, specially designed prompts, Few-Shot Learning, and dynamic truth discovery to extract exact human casualty statistics from crowdsourced text data with complex semantics, without additional training or fine-tuning. To the best of our knowledge, this is the first disaster human fatality information retrieval framework built on LLMs.
- (2) We design a physical constraint-aware dynamic truth discovery scheme to accurately uncover reported fatalities from noisy, incomplete, and conflicting information by considering (i) physical rules that human losses will not decrease with time, and (ii) historical reliability of different information sources.
- (3) We integrate the data pipeline, information extraction, and truth discovery with existing PAGER fatality loss models and enable automatic updating of the PAGER system in near-real-time seismic loss projection, for the first time.
- (4) We evaluate and characterize the framework using three recent real-world earthquakes. The evaluation results demonstrate significant performance gain achieved by our framework, providing timely and accurate human fatality information with finer time resolution compared to traditional approaches.

2 RELATED WORK

Although various near-real-time disaster information platforms are open-source for disaster response, there is still no framework openly available to support automatic and multi-source information

retrieval and impact estimation in near real-time. The USGS PAGER system provides rapid estimates of economic losses and human fatalities [18] relying on empirical models and geospatial data, and it can be updated by manually searching news sources for casualty reports [22]. In the Natural Language Processing (NLP) community, many approaches have been developed to acquire disaster damage information from social media platforms like Twitter. Some studies design pre-defined keyword lists or tables to search and extract useful information from social media textual posts. After creating keyword lists for each subcategory, the keyword search-based method can identify and categorize qualified posts and contain comprehensive situational knowledge[13]. For example, [8, 13] split information on disaster damage into multiple groups like infrastructure destruction, supply chain demands, and affected activities. Generally, methods of keyword searching usually discover certain information from the social media text corpus. After keyword lists are created for each group, this approach can identify and categorize qualified posts. However, it is costly and time-consuming to enumerate every possible keyword and phrase related to a topic due to the colloquial nature of social media textual messages [16, 28]. Previous work like [16, 27, 31] apply existing word embeddings or train disaster social-media-post-specific word embeddings to obtain social media data representations. Afterward, machine learning methods like linear classifiers, logistic regression, and Support Vector Machines (SVMs) are fine-tuned upon the embeddings of the datasets to recognize and categorize Twitter messages. Due to the excellent performance in image classification, Convolutional Neural Networks (CNNs) are deployed extensively [1, 2, 6, 21]. However, these methods are often ineffective when applied to unobserved events and need fine-tuning.

Transformer-based language models have recently become state-of-the-art due to their powerful attention mechanism that models inter-token relations [29]. The transformer models are usually trained on large amounts of online texts, making them applicable to many language tasks. Afterward, a Bidirectional Encoder Representations from Transformers (BERT) model, an Encoder-only variant of Transformers that outputs word and sentence-level representations, is applied as proposed by [9]. BERT outperforms traditional word embedding methods in many natural language understanding tasks by providing context-aware representations. In this study, we use notable BERT variants, RoBERTa, and XLM-RoBERTa [20].

Dynamic Truth Discovery: The truth discovery problem was first formally formulated and resolved by a Bayesian heuristic algorithm, Truth Finder, in [34]. Given estimated source weights and interactions among different claims, the confidence score of each claim is updated using Bayesian updating. Based on Truth Finder, extended models were further introduced to integrate prior knowledge, including constraints on truth patterns and source dependencies, to improve the accuracy and efficiency of truth discovery [10, 23]. However, significant knowledge gaps exist in finding truth from widely spread disingenuous posts and information with severe time latency under dynamically changing truths, especially for time-sensitive tasks like ours, which is a challenging task that yet has not been well addressed.

3 FRAMEWORK DESIGN

In this section, we present our framework (shown in Figure 1) to extract casualty data from crowdsourced reports. Specifically, this framework includes our key components: (1) an automatic data crawling pipeline that automatically scrapes data from multiple sources, (2) a hierarchical human cost value extraction module integrating a hierarchical event classifier that filters out text relevant to target earthquake events and casualty statistics (Section 3.1), and a fatality value extractor built based on LLMs and Few-Shot Learning (Section 3.2), (3) a physical constraint-aware dynamic truth discovery model that recovers casualty estimates from massive noisy and potentially conflicting data, constrained by physical rules of evolving reported fatalities (Section 3.3), and (4) a PAGER loss-projection model that dynamically updates final human cost estimations (Section 3.4). To enable near real-time disaster data retrieval, we build an event-triggering pipeline that retrieves and processes real-time disaster data, mainly text, from Twitter and News API. The pipeline enables automatic keywords and query generation as well as streaming data collection and storage. We mainly focus on extracting human casualty statistics from social media posts and news headlines and articles.

3.1 Hierarchical Event Classifier

Hierarchical event classifiers filter out irrelevant text data from large amounts of crowdsourced data to improve the computational efficiency of quantitative human loss data pairing. Our hierarchical event classifier contains two modules: a earthquake event classifier to tell if a text is relevant to a target earthquake event, and a fatality statistics classifier to determines if the text includes casualty statistics. The design of the integrative hierarchical event classifier is based on our observations concerning disaster text data. Two common phenomena we discovered are that (1) because disaster zones are often large in extent, there are often fatality reports therein that are not induced by the event of interest, but rather by unrelated occurrences (e.g., car accidents or pandemic); and (2) because seismic impacts are often complex, a large amount of earthquake-related information does not contain casualty statistics. Based on the observations, we design the hierarchical event classifier that cross-classifies the input text data to cull irrelevant information.

To deal with complex semantics in multi-lingual, multi-sourced data, the two modules share the same model architecture back-bone as XLM-RoBERTa, a state-of-the-art, cross-lingual word-embedding model and contains 350 million parameters [7] for effective word embedding. The word representations are further input to a neural network to classify if a text is relevant to an earthquake event as well as if the text contains any casualty statistics. XLM-RoBERTa was pre-trained on text spanning 100 languages, giving it multilingual understanding and cross-lingual transfer. The cross-lingual transfer capability enables us to only train the models with the abundant English language and generalize our classification to more resource-scarce languages. We further train the earthquake classifier and fatality statistics classifier separately using labeled disaster corpus, CrisisNLP [17]. CrisisNLP labels them through crowd-sourcing efforts. The social media posts enclose various disaster events (e.g., earthquakes, hurricanes, pandemics), labels that describe whether a social media post is relevant to the disaster, and statistics.

3.2 Human Loss Extraction via LLMs

With most irrelevant information filtered out by the hierarchical event classifier, we further extract the casualty numbers. The human casualty extractor plays two important roles: (1) to extract the exact number of casualties, including fatalities, injuries, and locations, and (2) to provide second verification of the relevance of the text data, e.g., past earthquake occurrences recalled in the same region, which can not be directly differentiated by the event classifier. To our best knowledge, ours is the first study to accomplish near-real-time casualty value extraction on crowdsourced text data other than the classification of messages. The desired details are often embedded within crowdsourced text data with complex language. For example, issues such as irregular syntax, use of the conjunctions 'and' or 'or,' abbreviations, and confusing numerical expressions need to be addressed. For example, when searching the 2021 M7.2 Haiti earthquake in Twitter, a relevant post is as follows:

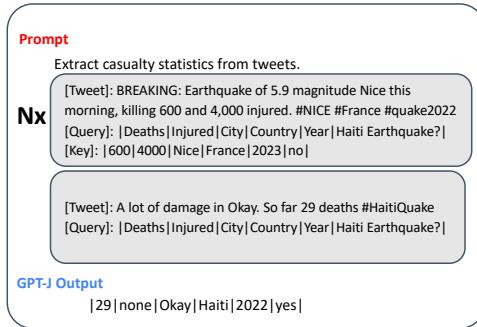


Figure 2: A conceptual diagram of our Few-Shot Learning prompt approach to extracting information. Nx represents the number of examples (Shots) that we give in the prompt.

"8/21 Haiti was hit by an earthquake leaving 2,200 dead, 10K homeless. 1 week later a Hurricane, killing 14, caused 500mil in damage. 1 month b4 they're Pres was killed leaving the isle lawless. Those are refugees fleeing death & devastation, they have nothing left to go back to."

This tweet contains multiple quantitative values related to the targeted earthquake (2,200 dead and 10,000 homeless), but also irrelevant information about a hurricane one week later than the earthquake (14 death, 500 million damage). Moreover, many multi-lingual abbreviations, such as local city names, cannot be directly filtered out by traditional rule-based methods. For example, we found this post for the aforementioned Haiti earthquake:

"A lot of damage in Okay. So far 29 reported deaths."

"Okay" is Haitian Creole for Les Cayes, a major port and city in Haiti that suffered severe damage during the 2021 Haiti earthquake. Because our system targets global earthquakes, it is impossible for the traditional natural language processing techniques—such as imposing manually designed rule-based or keyword matching—to handle such highly flexible text data reporting patterns varying with local social-cultural characteristics.

To address these challenges, we hereby use the new generation of LLMs and Few-Shot Learning to conduct unstructured data pairing for accurate and efficient casualty data extraction from crowdsourced text data.

3.2.1 Backbone LLMs. Among the commonly used transformer-based language models, the Generative Pre-trained Transformer (GPT), introduced by [24], achieves the most robust text generation. The GPT is an autoregressive model that uses seed text as context to generate new text [5, 25]. GPT-2 (1.5 billion parameters) [25] and GPT-3 (175 billion parameters) [5] are developed to enhance the capability of the model to recognize patterns present within the input text, without the need for fine-tuning using fully labeled datasets. In this work, we utilize GPT-J, an open-source alternative to the GPT family [30]. Although GPT-J only has 6 billion parameters, the model sufficiently retains the capabilities and embedded knowledge present within the comparably larger GPT-3 (175 billion).

Because casualty estimation frameworks like the PAGER system are particularly sensitive, the LLM doubles as a second layer of defense. We can eliminate distracting information by utilizing extracted information such as location and earthquake specificity. If we extract a statistic but cannot convert it to a number, we will discard it. Although this method is generally reliable, generative models can still produce random or inexact answers. These errors come from close numbers, random characters, and related words. To combat this, we run a beam search for the most likely response at inference. We further assure the data reliability by tracking the uncertainties of each produced token and limiting the probabilities to a specific range.

3.2.2 Prompt Design. Few-Shot Learning in natural language processing mainly refers to the practice of feeding a pre-trained language model with a very small number of natural language templates, i.e., "**prompt**," as opposed to fine-tuning methods that require a large amount of training data [26]. It helps the model adapt to the desired task with decent accuracy. This technique enables the model to generalize to understand related but previously unseen tasks with just a few examples. One of the key elements of Few-Shot Learning is prompt design. In our prompt design, we attach data to each text and ask the model to replicate the examples. Each example contains one or two sentences of text, a query, and the responses to the queries as follows:

[Tweet]: BREAKING: Earthquake of 5.9 magnitude in Nice this morning, killing 600 and 4k injured. #France#NICE
[Query]:|Deaths|Injuries|Location|Cities|Country|Year|Haiti Earthquake?
[Key]:|600|4000|Nice|France|2021|No|

We include examples that cover possible edge cases and missing information for increased robustness. For instance, some example text data do not contain injury statistics, and we will replace the response with a unique character that designates it as missing. We encourage the model to fill in incomplete or obfuscated information like location with its pre-existing knowledge base from the LLMs (e.g., recognizing that Okay is Haitian Creole for Les Cayes and that it is a city in Haiti).

3.3 Dynamic Truth Discovery

The truth discovery problem was first introduced to find the true claim from multiple claims shared by different information resources [34]. Researchers proposed multiple models (e.g., AVGLog, Invest, and PooledInvest) to handle the source dependency and heterogeneous source credibility in truth discovery problems [10, 23]. However, finding the truth when many posts are disingenuous and ground truth dynamically changes is still a challenging task. In this work, dynamic truth discovery is designed to integrate multiple

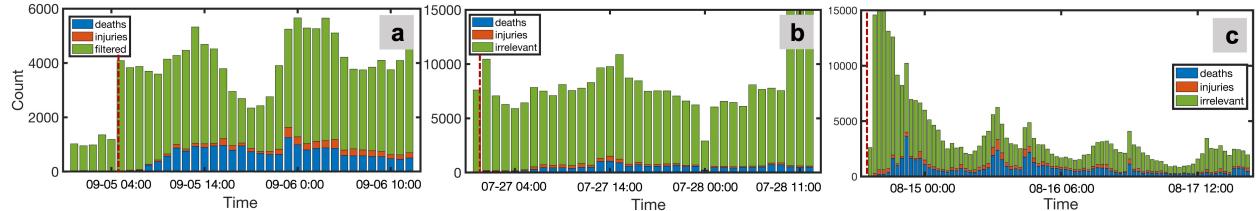


Figure 3: Distributions of social media posts filtered for relevance, posts mentioning fatalities, and posts mentioning injuries obtained for (a) the 2022 Luding, China earthquake, (b) the 2022 Philippines earthquake, (a) the 2021 Haiti earthquake as time evolves (UTC time). The red line represents when the mainshock occurred.

different information sources to yield a distribution of casualty values $p_t = (p_t^1, \dots, p_t^k, \dots, p_t^K)$. To explicitly model the quality of estimations from different sources, we design an information score ($IS_{i,t}^k$) to quantify the contributions of each data source i to the belief of human cost value k at a certain time point t . The information score is designed based on three aspects of the information credibility from a specific data source i : confidence score ($\xi_{i,t}^{u,k}$), relevance score ($r_{i,t}^{u,k}$), and independence score ($\rho_{i,t}^{u,k}$):

Confidence score ($\xi_{i,t}^{u,k}$) quantifies the confidence level of the extracted human cost variable k from a text data point u provided by the source i , with a range of $(0, 1)$. This score can be obtained from the confidence level of a large language model when answering a fatality query.

Relevance score ($r_{i,t}^{u,k}$) measures if a text data u output value of k is relevant to casualty information in the target disaster event, with a range of $(0, 1)$. Relevance is obtained by the probability output from the hierarchical event classifier and the probability of LLM's answers to query questions about the event.

Independence score ($\rho_{i,t}^{u,k}$) depicts if a text data u indicating casualties k is original data instead of forwarding/copying information from other earlier text data, with a range of $(0, 1)$. This score is obtained based on if a post is significantly similar to an earlier post or cites information from another source. Higher scores mean that the data source is more independent. Integrating the above scores, we define the information score as

$$IS_{i,t}^k = \sum_u \xi_{i,t}^{u,k} * r_{i,t}^{u,k} * \rho_{i,t}^{u,k}.$$

By normalizing the score across multiple individual accounts $i \in I$ to a range of $[0, 1]$, we get a normalized information score $NIS_{i,t}^k$. We also impose physical constraints to further calibrate the information score. The physical constraint is based on order statistics, i.e., *that fatality numbers should not decrease with time*. Therefore, the transitions from p_{t-1} to p_t should be subject to a constrained transition matrix, i.e., an upper triangular matrix, due to the probability of transiting from value m to any $n < m$ is zero. We can further obtain a hard upper bound for each value k 's probability, where $k \leq K$, at time point t :

$$NIS_{i,t}^k \leq p_t^k \leq \max(p_{t-1}^1, \dots, p_{t-1}^k).$$

For example, if at time point $t-1$, the probability of fatalities value 0 is 0, then the probability that it will become 0 at time point t is 0, because the number of deaths will only stay the same or increase. To impose this physical constraint, we will prohibit the invalid transition by removing the corresponding $IS_{i,t}^k$.

We also design a source reliability score (s_i) to quantify the reliability of the information provided by the source i . We denote source i 's output set as $g(i)$ and the set of sources that can output value k as $f(k)$. The reliability score is measured based on historical information credibility by summing up all the information scores of the source i . We apply a sigmoid function to normalize its scale between 0 and 1 and get D_t^k .

$$s_i = \frac{\sum_{k \in g(i), t} I(IS_{i,t}^k) D_t^k + (1 - I(IS_{i,t}^k))(1 - D_t^k)}{\sum_{k \in g(i), t} |IS_{i,t}^k|}, \quad (1)$$

where $D_t^k = \frac{1}{1 + \exp(-\sum_{i \in f(k)} IS_{i,t}^k)}$. $I(x)$ is an indicator function in which $I(x) = 1$ when $x > 0$ or else it is 0. The reliability score evaluates the ability of a source to provide high-fidelity estimates agreed by other high-fidelity sources. We finally obtain the updated probability distribution of the values as

$$p_t^k = \frac{\sum_{i \in g(k)} s_i NIS_{i,t}^k}{\sum_{i \in g(k), k \in f(i)} s_i NIS_{i,t}^k}. \quad (2)$$

To ensure the physical constraints persist, we will take the upper bound of p_t^k if the value is higher than the upper bound. By fusing different sources, the aggregated estimate at time t is:

$$k_t^* = \arg \max_k p_t^k. \quad (3)$$

The physical constraints-aware, dynamic truth discovery scheme finally outputs casualty values hourly, mainly deaths and injuries for the target event in each target country. We further use the timestamp of the first text reporting the corresponding human cost value, as the corresponding time label to obtain a reliable time-series of casualties devoid of any duplications or redundancy.

3.4 Fatality Estimate Projections

Past studies show that reported losses for many earthquakes follow a simple (but initially unpredictable) exponential cumulative distribution function, determined by parameter α . The loss projection model can be formulated as

$$N(t) = N_\infty (1 - \exp(-\alpha t)) \quad (4)$$

With fatalities reports extracted from crowdsourced data, we can update our estimates of the parameter α utilizing Bayesian updating. In this work, we follow the Bayesian updating algorithm used by the current PAGER system [22] to enable efficient fatality projection updating. The approach integrates the uncertainties of new observations from reported data with the *a priori* model learned from historical events occurring in similar regions. Currently, due to the significant impacts on the PAGER system results, USGS experts still need to carefully review and validate the aggregated fatality estimates extracted from dynamic truth discovery before

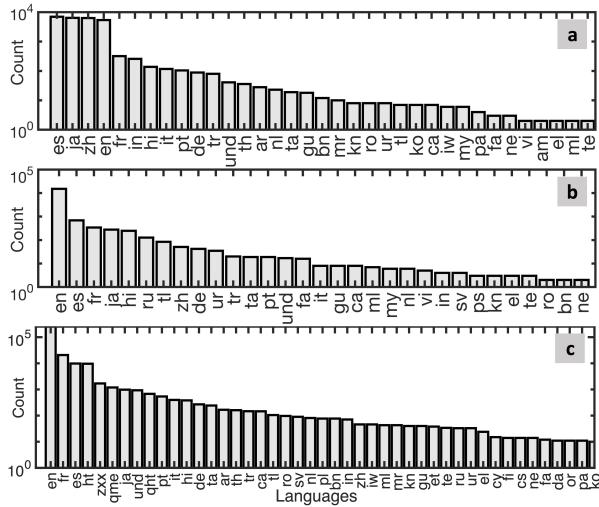


Figure 4: Language distribution of Twitter data retrieved for different earthquake events: (a) the 2022 Luding, China earthquake, (b) the 2022 Philippines earthquake, (c) the 2021 Haiti earthquake.

integrating them for human loss projection. In the future, the proposed framework is expected to further reduce the workload of 24x7 on-call experts as the LLMs and truth discovery algorithms improve.

4 RESULTS

Our framework has been fully deployed in real-time testing to provide global earthquake event information to PAGER system for more than half a year. We have also tested our framework on a sequence of significant recent earthquake events, each denoted by a magnitude (M) on the Moment Magnitude Scale, which measures the total energy released by an earthquake. The events include the M7.2 Haiti (2021), the M7.0 Luzon, Philippines (2022), the M6.8 Luding/Sichuan, China (2022), the M6.5 Taiwan (2022), the M6.8 Michoacan, Mexico (2022), the M7.6 Papua New Guinea (2022), the M5.7 Khowy, Iran (2022), and the M5.6 Indonesia (2022) earthquakes.

Here we present an evaluation of the performance of our framework on these real-world events. First, we present our experimental setup and performance evaluation metrics for the framework. We then characterize the data retrieved by the data pipeline, after the hierarchical event classifier, and after casualty value extraction. We also characterize the accuracy and error of the hierarchical event classifier. Finally, we evaluate the casualty estimation performance of our framework. The experimental evaluations are based on three aforementioned earthquakes: the 2021 Haiti earthquake, the 2022 Philippines earthquake, and the 2022 Luding, China earthquake, which caused substantial damage and fatalities.

Experimental Setup: The framework is triggered based on the magnitude of earthquake events. Our data sources include news data from News API and social media data from Twitter API (Academic research account). Each News API call retrieves a maximum of 100 records every half hour. Each Twitter API call retrieves up to 10,000 tweets every half hour. Twitter data provide timestamps, tweet content, user profiles, geotags, relevant news and images links, device

type, and other metrics. News data are retrieved from News API (<https://newsapi.org/>), which covers numerous news sources and media in 14 languages from 55 countries. The fatality data extraction process (including the backbone LLMs) is implemented using PyTorch v1.12.0 and the Docker system and conducted on a server with four NVIDIA RTX A6000 Graphics Processing Units (GPUs). In the real-time process, a half-hourly text data batch is fed into the information extraction model and automatically outputs deaths, injuries, city, country, if the tweet is relevant to the earthquake of interest, and which year the event occurs, as well as confidence scores associated with these answers. The results are automatically saved for dynamic truth discovery. In the truth discovery phase, we integrate all data points since the last time the fatality values are updated and finally provide the latest values as well as the first time that that value appears in the scraped text data. Finally, the casualty output data points can be fed into the PAGER loss model to update the overall fatality estimates for the earthquake.

Performance Evaluation Metrics: We characterize and evaluate our framework from two perspectives: timeliness and accuracy. Note that the final goal of this framework is to automate the fatality information extraction process to reduce the 24x7 operations expert's workload and improve the accuracy of loss estimates. To evaluate the timeliness, our goal is to achieve better or similar timeliness as manually retrieved data, for example, extracting the same casualty values earlier than manual extraction. Manually searching for casualty values is very time-consuming, requires personnel available at all hours, and depends on the agility of the expert with a wide range of search tools and social media platforms. The accuracy is two-fold: text classification accuracy and fatality number accuracy. Because it is impossible to label every text data retrieved in our real-world experiments, we evaluate the hierarchical event classifier mainly utilizing CrisisNLP data using the accuracy rate, F1 score, and false positive rate (FPR). The FPR represents the percentage of irrelevant texts that are classified as relevant and passed to the fatality value extraction. An ideal event classification model needs to be accurate and minimize false-positive cases, as fatality information is often sensitive and critical. Moreover, to evaluate the accuracy of the final extracted fatality value, we compare it with the officially released fatality number (often after weeks of an event).

Retrieved Data Overview: The original data retrieved from news and social media platforms mainly include a variety of languages and sources. For example, in the Luding, China, earthquake, most of the retrieved texts used Japanese, Chinese, Spanish, and English, as shown in Figure 4(a). Whereas for the Philippines earthquake, most texts used English, Spanish, French, Japanese, Hindi, and Filipino as shown in Figure 4(b). Haiti earthquake data contain more diverse languages, dominated by English, French, Spanish, and Haitian Creole (Figure 4(c)). The distribution of languages also depicts who shares and cares about disasters or disaster-related information, combined with the Twitter account profile. We found that the majority of data are from the affected zone. In the meantime, social media accounts from earthquake-prone countries such as Japan or from neighboring countries are also actively forwarding human fatality-related information. The long-tail effect is more prominent in the Philippines event compared to that in China. This observation also helps explain why traditional methods that only focus on English

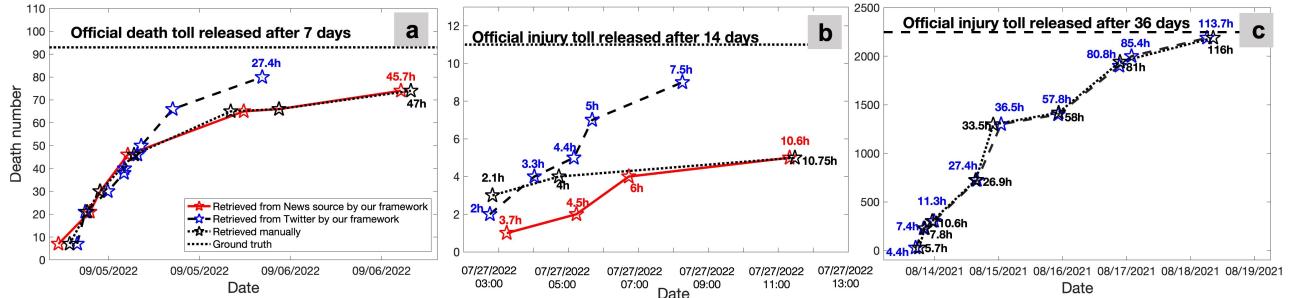


Figure 5: A comparison between fatalities extracted from Twitter, news articles, and manually searched for (a) the 2022 Luding, China earthquake, (b) the 2022 Philippines earthquake, (c) the 2021 Haiti earthquake as time evolves (UTC time). The solid red line represents news-sourced data and the dashed line with blue markers refers to Twitter-sourced data. The dashed line with black markers represents those manually searched. The black dotted line without any marker indicates the final official toll. Each updated fatality data point is labeled with the earliest reporting time.

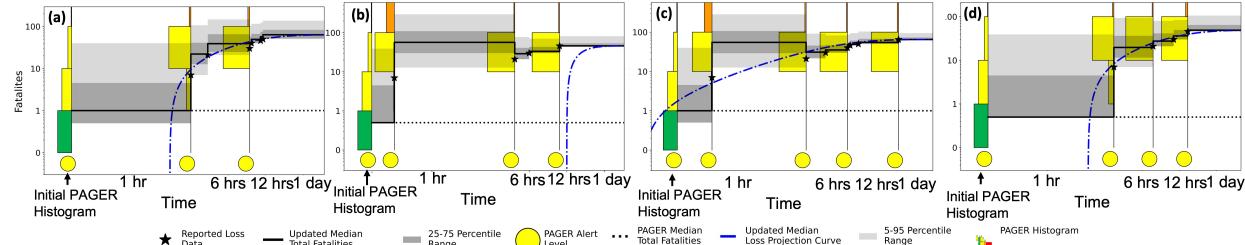


Figure 6: PAGER fatality estimate updating using (a) Twitter data, (b) news data, (c) mixed data of Twitter and news, and (d) manually searched data

text alone may not be generalized to predict disaster impacts for global earthquake events. Moreover, we found that the number of relevant texts containing casualties often increases quickly within a few hours of the event, and gradually converges as sufficient resources are allocated to the scene, as shown in Figure 3(a).

Classified Event-Casualty-Related Text Information: We evaluate the performance of our hierarchical event classifier. Our RoBERTa-based models predict accuracies of 97.4%. As for XLM-RoBERTa, we obtain a classification accuracy of 96.7%, a false positive rate of 0.050, casualty statistics classification accuracy of 96.1%, and a FPR of 0.045. Although the performance slightly dips, the multilingual benefits of the XLM-RoBERTa model. In total, there are 388 test samples for the statistic classifier and 3033 test samples for the earthquake classifier. The results show that, beyond the FPR, the two classifiers are separate and work together to effectively filter out the majority of irrelevant tweets.

We also visualize the distributions of death-related tweets, injury-related tweets, and irrelevant tweets for the three earthquakes in Figure 3. A common pattern that can be observed is that the number of related tweets increases quickly after an earthquake event occurs and gradually reduces. The rate of reduction is related to the actual death number – usually if an earthquake cause severe human fatality, such as the Haiti earthquake that causes thousands of deaths, the number of human fatalities will be kept updated in social media for a long time (more than 5 days). Meanwhile, if the human fatality number is not significant, such as in the Philippines earthquake in 2022, the number of human fatality-related tweets shrinks quickly, as Figure 3(b) shows.

Table 1: Results of different backbone LLMs for extracting death tolls from the Twitter platform for the Luding, China earthquake, compared to manual search in News platform (NA means no corresponding death number is extracted).

Deaths	Time since earthquake occurs (h)			
	GPT-J (6B params)	GPT-Neo (1.3B params)	BERT	Manual Search
7	3.0	3.1	3.1	2
21	4.1	6.7	5.9	4.3
30	7.0	9.2	8.9	6
38	9.2	9.3	NA	NA
40	9.2	NA	10.7	NA
46	10.9	11.0	11.0	10.5
50	11.4	NA	19.2	NA
66	15.6	26.1	31.1	29.6

Human Fatality Estimates: With text data filtered by a hierarchical event classifier, we further extract the exact number of human fatality information using LLMs and dynamic truth discovery. In this section,

We also analyze the results of human fatality information extraction and human fatality forecasting based on the extracted human fatality information on three major earthquake events. Due to limited space, we mainly show results for the 2022 M6.8 Luding, China Earthquake, the 2022 M7.0 Philippines Earthquake, and the 2021 M7.2 Haiti Earthquake, and summarized the results in Figure 3 and 5. On 04:52 September 5, 2022, UTC time, an *M* 6.8 earthquake struck Luding County, Sichuan Province, in southwest China. A

national earthquake emergency response (Level 3) was immediately launched by the Ministry of Emergency Management of the People's Republic of China and then upgraded to Level 2 on September 6. Based on extensive field surveys, an intensity map was provided by the Ministry of Emergency Management on September 11 [4]. Our data collection pipeline was triggered soon after the earthquake to collect crowdsourced data from Twitter and through News API. Using our framework, the death and injury numbers were extracted as input to the PAGER loss updating the platform. We compared the timeliness of our retrieved data from the News API, Twitter, and manual search in Figure 5(c). It can be seen that Twitter data are updated more than either the news or our manual search data. Especially, the death numbers of 66 and 80 are extracted 14 hours and 20 hours earlier than the manual search. We also compare the results of LLMs with different capacities in Table 1. It can be seen that GPT-J with 6 billion parameters more closely matches the manual search than GPT-Neo with 1.3 billion parameters and BERT. Currently, considering the limited bandwidth of deploying LLMs, GPT-J model presents a competitive capability of extracting information effectively. Meanwhile, as model parameters increase, a more powerful GPT or Open Pre-trained Transformers (OPT) model may further substantially improve our information extraction performance. Moreover, we utilize the data obtained from Twitter, news, and mixed data of both sources to update the PAGER loss estimation models, compared to the loss estimation performance using manually searched data, as shown in Figure 6. The figures present how the forecasted probability distribution of final human casualty is updated as new data points come in. It can be seen that mixed data and news results as well as news results alone provide an estimation that the final death number will fall into the range between 10 and 100 – which is later verified to be 93 deaths – earliest compared to Twitter and manual search. All four types of methods provide correct forecasting because the first data point is received, demonstrating that the human fatality information retrieved by our framework can achieve comparable performance compared to manual search by human experts.

The 2022 Philippines Earthquake: The M7.0 earthquake struck the northern Philippines caused 11 deaths and 615 injuries. The recent occurrence of the earthquake makes it an ideal opportunity to experiment with actual, real-time performance of our model. Due to the large Tagalog and Filipino-speaking populations, we apply our XLM-RoBERTa-based hierarchical event classifier. Our model reports that the number of deaths will increase from 1 to 10, and the number of injuries will increase from 44 to 60 over time following the earthquake. Likewise, the official death toll was released a week later, which fell outside our time frame. To benchmark our social media findings, we attempted to exploit news data for official casualty reports in the 2022 Philippines earthquake. For the news data, we treat each description as short text as a tweet, and process them similarly. As shown in Figure 5(b), we see a notable gap between the Twitter-sourced data and many echoing data points when we source our news feeds. This gap may come from the slower speed that official outlets have compared to social media outlets. The M7.2 earthquake struck Haiti result in a total number of 2,248 deaths and 12,200 injured. We crawled the Twitter database for tweets containing relevant keywords or hashtags (e.g., earthquake, Haiti) to obtain the social media data for that event. We process each

tweet with our method and obtain a time-series graph shown in Figure 5(c). Throughout the duration, our extracted death toll rises from 29 to 2,189. We observe that our extract statistics are relatively close to the final official number but still have a slight difference because of the limited time span of the deployment.

5 CONCLUSION

This paper presents a novel framework for near-real-time, earthquake-induced casualty estimation from multilingual social media data. We introduce a hierarchical event classifier that categorizes and filters informative social media posts with multilingual capabilities, a process previously unexplored. We extend this casualty data extraction beyond simple categorization and directly extract relevant statistics from the tweets, including locations and uncertainties. We overcome the challenges of complex syntax and requirement-labeled data in real-time direct extraction through large language models, leveraging its capabilities of Few-Shot Learning and LLMs. We design a physical constraint-aware dynamic truth discovery model that recovers causality estimates from massive noisy and conflicting data. In our experiments, we measure the capacity of our classification networks and evaluate the performance of our model on real-world events. Our results demonstrate that our model yield results that compare well with final reported losses and accurately extracted information automatically to significantly improve the timeliness and efficiency of the existing USGS PAGER system.

ACKNOWLEDGMENTS

This draft manuscript is distributed solely for informational purposes. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy. Any mention of commercial products is for informational purposes and does not constitute an endorsement by the U.S. government.

REFERENCES

- [1] Sajjad Ahadzadeh and Mohammad Reza Malek. 2021. Earthquake damage assessment based on user generated data in social networks. *Sustainability* 13, 9 (2021), 4814.
- [2] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain adaptation with adversarial training and araph embeddings. *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- [3] Reem ALRashdi and Simon O'Keefe. 2019. Deep learning and word embeddings for tweet classification for crisis response. *arXiv preprint arXiv:1903.11024* (2019).
- [4] Yanru An, Dun Wang, Qiang Ma, Yueren Xu, Yu Li, Yingying Zhang, Zhumei Liu, Chunmei Huang, Jinrong Su, Jilong Li, et al. 2022. Preliminary report of the 5 September 2022 MS 6.8 Luding earthquake, Sichuan, China. *Earthquake Research Advances* (2022), 100184.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
- [6] Cornelia Caragea, Nathan J McNeese, Anuj R Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, Andrea H Tapia, C Lee Giles, Bernard J Jansen, et al. 2011. Classifying text messages for the Haiti earthquake. In *ISCRAM Citeseer*.
- [7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).
- [8] Qing Deng, Yi Liu, Hui Zhang, Xiaolong Deng, and Yefeng Ma. 2016. A new crowdsourcing model to assess disaster using microblog data in typhoon Haiyan. *Natural Hazards* 84, 2 (2016), 1241–1256.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

- [10] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment* 2, 1 (2009), 550–561.
- [11] Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. 2011. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems* 26, 3 (2011), 10–14.
- [12] Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4, 1 (2013), 1–27.
- [13] Haiyan Hao and Yan Wang. 2020. Leveraging multimodal social media data for rapid disaster damage assessment. *International Journal of Disaster Risk Reduction* 51 (2020), 101760.
- [14] J Brian Houston, Joshua Hawthorne, Mildred F Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R Halliwell, Sarah E Turner McGowen, Rachel Davis, Shivani Vaid, Jonathan A McElderry, et al. 2015. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters* 39, 1 (2015), 1–22.
- [15] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47, 4 (2015), 1–38.
- [16] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 1021–1024.
- [17] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (Portoroz, Slovenia, 23–28). European Language Resources Association (ELRA), Paris, France.
- [18] Kishor S Jaiswal and David J Wald. 2010. Development of a semi-empirical loss model within the USGS Prompt Assessment of Global Earthquakes for Response (PAGER) System. In *Proceedings of the 9th US and 10th Canadian Conference on Earthquake Engineering: reaching beyond borders*. 25–29.
- [19] Stephanie Lackner. [n. d.]. *Earthquakes and Economic Growth*. <https://www.econstor.eu/bitstream/10419/194225/1/1043719490.pdf>.
- [20] 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* (2019).
- [21] Dat Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks. In *Proceedings of the international AAAI conference on web and social media*. <https://www.aaai.org/ojs/index.php/ICWSM/ICWSM17/paper/view/15655>
- [22] Hae Young Noh, Kishor S Jaiswal, Davis Engler, and David J Wald. 2020. An efficient Bayesian framework for updating PAGER loss estimates. *Earthquake Spectra* 36, 4 (2020), 1719–1742.
- [23] Jeff Pasternack and Dan Roth. 2010. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 877–885.
- [24] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
- [25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [26] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980* (2020).
- [27] Kevin Stowe, Michael Paul, Martha Palmer, Leysia Palen, and Kenneth M Anderson. 2016. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. 1–6.
- [28] Irina Temnikova, Carlos Castillo, and Sarah Vieweg. 2015. EMTerms 1.0: A Terminological Resource for Crisis Tweets. In *ISCRAM*.
- [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
- [30] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>.
- [31] Yandong Wang, Shisi Ruan, Teng Wang, and Mengling Qiao. 2019. Rapid estimation of an earthquake impact area using a spatial logistic growth model based on social media data. *International Journal of Digital Earth* 12, 11 (2019), 1265–1284.
- [32] Zheye Wang and Xinyue Ye. 2018. Social media analytics for natural disaster management. *International Journal of Geographical Information Science* 32, 1 (2018), 49–72.
- [33] Max Wyss. 2017. Report estimated quake death tolls to save lives. *Nature* 545, 7653 (2017), 151–153.
- [34] Xiaoxin Yin, Jiawei Han, and Philip S Yu. 2007. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*. 1048–1052.

SUPPLEMENT

5.1 Additional analysis for the 2022 Luding, Sichuan Earthquake

Figure 7 presents the independent score, confidence score, and relevance score defined in Section 3.3 to weight different source information for human fatality statistics extraction in the 2022 Luding, China earthquake. It can be seen that comparing the verified account (True) and unverified account(False), the independence score of verified account is relatively higher than unverified ones, showing that they are more independent in tweeting the information compared to unverified accounts which are mostly personal users. Similarly, the confidence score and relevance score of verified accounts are also higher than unverified ones. We also present Figure 8 to show the distributions of extracted death number as time evolves, as well as their mode. It can be seen that as time changes, the modes of extracted human death number increases from 7 to 46 within 9 hours. Our truth discovery algorithm aggregate these extracted death number more effectively to discover the truth earlier than simply taking the mode. For example, the death number of 21 first appears as a mode in 7 hours after the earthquake, while our algorithm discovers 21 as death number in 4.1 hours after the earthquake occurs. The results show the effectiveness and timeliness of our dynamic truth discovery combining with large language models.

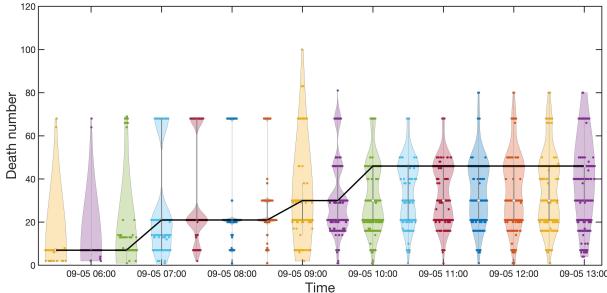


Figure 8: Distributions of extracted human death number as time evolves.

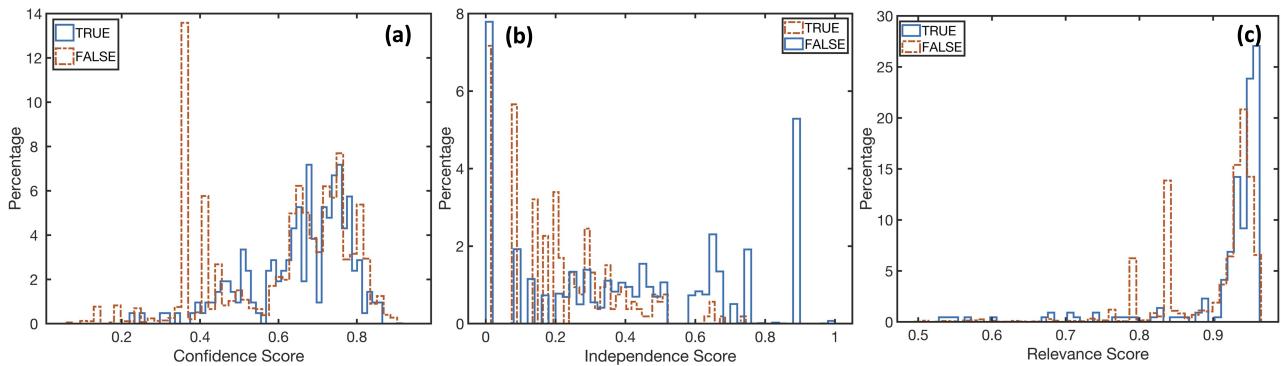


Figure 7: Confidence score, independence score, and relevance score for verified Twitter account (true, blue rectangle) and unverified Twitter account (false, orange dashed rectangle) for the filtered death tweets in the 2022 Luding, China earthquake.

Model	Acc (%)	F1	FP Rate
RoBERTa EC 3-Epochs	97.2	0.97	0.042
RoBERTa EC 4-Epochs	97.4	0.97	0.034
XLM-RoBERTa EC 3-Epochs	96.7	0.97	0.050
RoBERTa SC 3-Epochs	95.6	0.96	0.045
RoBERTa SC 4-Epochs	95.9	0.96	0.061
XLM-RoBERTa SC 4-Epochs	96.1	0.96	0.045

Table 2: Performance of the hierarchical event classifier, which integrates an earthquake classifier (EC) and a human cost statistics classifier (SC), both based on RoBERTa/XLM-RoBERTa.