Dec 11th, 12:00 AM

Transformer-Based Multi-Task Learning for Crisis Actionability Extraction

Yuhao Zhang
Singapore Management University, yuhaozhang@smu.edu.sg

Siaw Ling Lo
Singapore Management University, sllo@smu.edu.sg

Phyo Yi Win Myint
Singapore Management University, ywmphyo@smu.edu.sg

Transformer-Based Multi-Task Learning for Crisis Actionability Extraction

Completed Research Paper

Yuhao Zhang
Singapore Management University
80 Stamford Rd, Singapore 178902
yuhaozhang@smu.edu.sg

Siaw Ling Lo
Singapore Management University
80 Stamford Rd, Singapore 178902
sllo@smu.edu.sg

Phyo Yi Win Myint
Singapore Management University
80 Stamford Rd, Singapore 178902
ywmphyo@smu.edu.sg

Abstract

Social media has become a valuable information source for crisis informatics. While various methods were proposed to extract relevant information during a crisis, their adoption by field practitioners remains low. In recent fieldwork, actionable information was identified as the primary information need for crisis responders and a key component in bridging the significant gap in existing crisis management tools. In this paper, we proposed a Crisis Actionability Extraction System for filtering, classification, phrase extraction, severity estimation, localization, and aggregation of actionable information altogether. We examined the effectiveness of transformer-based LSTM-CRF architecture in Twitter-related sequence tagging tasks and simultaneously extracted actionable information such as situational details and crisis impact via Multi-Task Learning. We demonstrated the system's practical value in a case study of a real-world crisis and showed its effectiveness in aiding crisis responders with making well-informed decisions, mitigating risks, and navigating the complexities of the crisis.

Keywords: Actionability, Crisis Response, Multi-Task Learning

Introduction

Crisis informatics studies the use of information and technology in crisis response. With the prevalence of social media platforms, the vast amount of social media data prompted researchers to study its potential role in crisis informatics. Traditional news media usually produce centralized, curated reports with structured formatting. Social media platforms, on the other hand, due to their rich information sources and crowdsourced reporting angles, tend to offer alternative information that may not be covered by their traditional counterparts. On the flip side, the decentralization of narratives on social media can cause opposing angles, conflicting information, false reporting, and rumors (Raza and Ding 2022; Gidwani and Rao 2023). The problem is further aggravated by the ambiguity of problem definition which leads to different decisions in what information should be considered relevant or useful. In the case where relevant messages were correctly identified, it is often unclear how they can be used to effectively aid crisis response. For example, the classification of crisis-related tweets, despite being one of the most studied tasks, has a low practical value in crisis response as the label of "relevant/irrelevant" alone does not provide useful

information. More often than not, crisis responders end up with an overwhelming number of relevant tweets with little actionable insight. The issue of cognitive load on emergency operation center personnel has not been effectively addressed. The inefficiency of current solutions and the lack of readiness for customization for the different needs of different stakeholders are among the key reasons for the low adoption rate by field practitioners (Reuter et al. 2018; Suwaileh et al. 2022).

It was believed that for the wider adoption of social media processing systems among crisis practitioners, the emphasis should be placed on identifying actionable information (Coche et al. 2021). Actionable information was defined as information that enabled an immediate action or could be acted upon by crisis responders. Recent fieldwork with crisis managers and operators (Zade et al. 2018; Kropczynski et al. 2018; Snyder et al. 2019; Hiltz et al. 2020) suggested that in crisis response, the primary information need was actionable information instead of general situational awareness. Identifying actionable information from social media during a crisis is crucial for effective crisis response and mitigation. The process of actionability extraction begins with the identification of actionable messages followed by more steps which may include crisis impact or severity estimation, locating affected areas, identifying key entities or urgent needs, information aggregation and visualization. While some systems have been developed to identify actionable information on social media during a crisis, they only completed the first step of actionability extraction, that is, the identification of actionable messages (Purohit et al. 2018; Snyder et al. 2019; Wang et al. 2021). They identified actionable tweets without providing details that made these tweets "actionable" in the first place. For example, a message calling for help to rescue an injured person was identified as actionable: "@USER: heard there's a riot going on now...police injured and cars set on fire! Pls help". But crisis responders were unable to conduct the evacuation without geolocation information, which may potentially be extracted from other tweets reported during the event, e.g., "Riot broke out at #littleindia. Crowd gathered at the scene". While some works focused on crisis location extraction (Hernandez-Suarez et al. 2019; Hu et al. 2020; Grace 2021), they did not extract other situational details or crisis impact. No systems that handle all steps of crisis actionability extraction have existed so far.

While the definition of actionable information was self-explanatory, various considerations existed during the implementation of actionability in a system. Coche et al. (2021) suggested that actionability should entail relevancy, timeliness, precision, and reliability. In this study, we considered two essential components: situational details and crisis impact. The situational details are the "what", "where", "who" and "when" of a crisis, which are enablers of a crisis dispatch (i.e., what to act upon). For instance, when a report of an injured person is received, it is crucial for the crisis responders to get hold of location information to conduct the evacuation. The crisis impact factors (e.g., affected individuals or infrastructure/utility damage) are used to understand the crisis severity and prepare crisis responders for resource allocation and prioritization which is critical given the limited resources in practical situations.

In this study, we implemented a transformer-based sequence tagging model via Multi-Task Learning (MTL) to simultaneously extract actionable phrases including situational details and crisis impact. We evaluated the effectiveness of BERTweet (Nguyen et al. 2020), the state-of-the-art transformer-based embedding for tweets, as well as the LSTM-CRF architecture for the aforementioned sequence tagging tasks on social media. In turn, we proposed a Crisis Actionability Extraction System (CAES) for social media which was the first social media processing system that handled filtering, classification, phrase extraction, severity estimation, localization, and aggregation of actionable information altogether. It specifically addressed actionability in urban emergency response, bridging a significant gap in existing crisis management tools. We demonstrated the practical value of CAES as a real-world application and its effectiveness in actionability extraction using a case study of a real-world urban emergency.

In the next section, we examined some related work. We explained sequence tagging, categorization choices, and MTL approach in the "Method" section. In the "Experiment" section, we discussed the dataset, model configuration, hyperparameter tuning, model training, and evaluation for the MTL model. In the "System Design" section, we explained the system workflow of CAES and the functionalities of individual modules required for actionability extraction. The "Case Study" section showed how CAES extracted actionable information to aid crisis responders in the crisis of 2013 Singapore Little India Riot. We also discussed the limitations and made suggestions for future work in "Challenges and Future Work" before conclusions were drawn in the last section.

Related Work

Information Extraction for Social Media Crisis Response

Various tasks have been proposed in social media crisis informatics, including but not limited to, detecting natural hazards (Poblete et al. 2018; Hernandez-Suarez et al. 2019; Krishnan et al. 2023), filtering relevant tweets (Mazloom et al. 2019; Ning et al. 2019; Snyder et al. 2019; Domala et al., 2020; Zahra et al. 2020; Alam et al. 2021; Lo et al. 2023), analyzing tweet sentiments (Muhammad et al. 2022; Rodrigues et al. 2022) and summarizing crisis information (Rudra et al. 2018; Ning et al. 2019; Ampili and Kanakala 2022; Garg et al. 2023). As the short, noisy characteristics of social media texts posed a major challenge in effective information extraction, researchers have employed word embeddings or Language Models (LM) specifically trained on crisis-related social media texts for better feature representation. Liu et al. (2021) proposed an attention-based, document-level contextual crisis embedding based on BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019). It outperformed conventional static word embeddings such as Word2Vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014) in crisis classification. Nguyen et al. (2020) introduced BERTweet, a BERT-like LM trained with English tweets using the RoBERTa (Liu et al. 2019) pre-training procedure. BERTweet outperformed previous state-of-the-art models on Twitter-related tasks such as part-of-speech (POS) tagging, named entity recognition (NER) and text classification. With the advancement in natural language processing (NLP), using a pretrained contextualized embedding such as BERT and BERTweet is now a de facto approach in almost all NLP applications. However, the effectiveness of BERTweet for information extraction on crisis-related tweets has rarely been studied. In this study, we compared BERTweet with the standard BERT as well as the GloVe embedding trained on tweets in extracting relevant information from crisis-related tweets.

While much effort has been made by technologists to utilize information on social media for crisis response, the current solutions are rarely deployed by field practitioners (Reuter et al. 2018; Suwaileh et al. 2022; Zhang et al. 2023). Several works have tried to bridge the gap between research and application by understanding the needs of relief organizations and the utility of existing solutions (McCreadie et al. 2019; Hiltz et al. 2020; McCreadie et al. 2020; Kruspe et al. 2021). Recent fieldwork suggested that the focus should shift from general situational awareness to actionability in order to improve the utility of social media data for crisis response (Kropczynski et al. 2018; Zade et al. 2018). Coche et al. (2021) made suggestions for the design of future crisis processing systems, among which was improving the identification of actionable information. While they defined actionability as "relevant, timely, precise, and reliable", others suggested that the definition should be user dependent. There were also several systems developed to identify actionable information on social media. Purohit et al. (2018) designed a message serviceability model, which tried to determine whether a message contained sufficient detail to identify a location, time, or related markers necessary to direct a response, thus actionable. While the system could identify serviceable requests by classifying and ranking them, it did not address the information extraction part of the actionability. In other words, the labels did not tell us what information made the particular message actionable. Snyder et al. (2019) proposed a system for identifying important, relevant data quickly through interactively training the model to remove noise and filtering by relevance. Again, the system did not further process relevant tweets for information extraction. This could lead to information overload on emergency operation center personnel. TRECS Incident Streams (TRECS-IS) (McCreadie et al. 2019; McCreadie et al. 2020) was an initiative designed for the categorization of crisis-related tweets and defined an ontology of 25 information types, among which 6 were defined as "actionable". For instance, when a user was reporting an emerging threat, the message was classified as "Report-EmergingThreats" and deemed "actionable". Based on TRECS-IS dataset, Wang et al. (2021) implemented a transformer-based MTL model for classifying information types and estimating the priority of tweets. However, the model only performed the classification task of identifying actionable tweets without labeling details. For example, when a message reporting an emerging threat was identified as "actionable", crisis responders would have to search for relevant details in the tweet manually. A second-level processing was necessary to extract actionable information automatically. A sequence tagging framework was proposed in our study where not only crisis-related tweets were filtered but also actionable information such as situational details (e.g., location, time, entities) and crisis impact (e.g., fatality, injury, property damage) were extracted from them directly.

McCreadie et al. (2020) suggested that future systems should integrate geolocation pipelines to extract explicit crisis location as it was recognized as an important component of actionability. Hernandez-Suarez

et al. (2019) developed a model to classify crisis impact and extracted crisis locations via NER. Other systems adopted a similar two-step approach which was classification of crisis-related tweets followed by geolocation extraction (Hu et al. 2020; Grace 2021; Suwaileh et al. 2022). As location entities were extracted from crisis-related tweets after the classification step, the relatedness of such locations to a crisis required verification, without which the reliability of such actionable information was questionable. Ning et al. (2019) built a CNN-based model with manually engineered features to identify informative tweets during crises with awareness of source types. A rule-based method using POS tags and occurrence frequency was applied to extract crisis and damage clues (relevant phrases from tweets). For example, any phrase starting with a number was considered a damage-related clue. The visualization of the proposed crisis relational topology was helpful in understanding the crisis dynamics by showing co-occurrence among certain key phrases. However, the accuracy of these extracted phrases was left unexamined. In addition, actionable knowledge was treated indiscriminately, without providing a clear categorization. In our study, we organized actionable information into six situational detail types and three crisis impact types (see section "Method"). Our end-to-end crisis actionability extraction system (CAES) handled filtering, classification, phrase extraction, severity estimation, localization, and aggregation of actionable information altogether.

Review of Sequence Tagging Models

As we proposed to extract actionable information via sequence tagging, we reviewed some existing deep learning-based sequence tagging models, which typically consisted of three components: embedding, context encoder and tag decoder (He et al. 2020; Li et al. 2023). The embedding module utilized a pretrained word embedding to map words into their distributed representations as the initial input of the model. Transformer-based models such as BERT and BERTweet can encode richer semantics and generally outperform static word embeddings (Lo et al. 2023). The context encoder extracted contextual features and dependencies of an input sequence and passed the learned features into the tag decoder for label prediction. Bidirectional LSTM (bi-LSTM) (Graves et al. 2013) was the predominant context encoder (Xia et al. 2019; Luo et al. 2020; Zhang et al. 2023) due to its effectiveness in representing the global information by incorporating contexts from both forward and backward directions. While Convolutional Neural Network (CNN) was another popular option for its efficiency, it had difficulties in capturing long-range dependencies. Lastly, the tag decoder took the hidden states from the context encoder as input and predicted sequence labels. Conditional Random Field (CRF) (Lafferty et al. 2001) considered the correlation between labels of adjacent tokens by computing transition scores, whereas softmax assumed that hidden states from the context encoder were conditionally independent, which might lead to label bias problem. CRF was thus widely used during tag inference among neural sequence tagging models (Al-Zaidy et al. 2019; Fan et al. 2019; Chen et al. 2023). Given the effectiveness and popularity of bi-LSTM-CRF architecture, we combined it with the BERTweet embedding to build the sequence tagging model in our study. One may question the necessity of a context encoder (i.e., LSTM) given that BERTweet already modelled contextual information. Given this valid concern, we explored by creating a BERT baseline model with a fully connected layer to assess the need for a context encoder in the presence of contextualized embeddings.

Method

Sequence Tagging and Annotation

A sequence tagging framework was proposed to extract tweet spans containing actionable information for crisis responders. A sequence tagging task takes a sequence of tokens $x = (x_1, \ldots x_n)$ as the input and predicts a label for each token. The output is a sequence $y = (y_1, \ldots y_n)$, where each y_i is the label of x_i. A consecutive sequence of tokens with the same label type is referred to as a span. Figure 1 shows two examples of crisis-related tweets during 2013 Singapore Little India Riot. In Figure 1(a), the underlined tokens "riot", "at", "Mustaffa", "Centre" were labelled with the same type "Civil Disorder". As the tokens were consecutive, the entire sequence "riot at Mustaffa Centre" was a span labelled as "Civil Disorder". Tokens without underline were labelled as non-entity tokens (commonly denoted as "O" in NER), that is, containing no actionable information. In the conventional classification framework, the entire tweet is given a label (e.g., crisis type or informativeness). However, these labels do not tell us what information contained within the particular tweet made it actionable. Additional steps are required to extract information which may be acted upon by crisis responders, e.g., named entities or geolocations (Jiang et al. 2022; Suwaileh et al. 2022). The sequence tagging framework, by comparison, can directly extract spans with this

information. In addition, irrelevant entities outside labelled spans are removed whereas in classification framework, all entities are candidates, and their relevancy to the crisis requires verification. Sequence tagging not only label actionable spans directly but also ensures that only relevant entities are extracted if further processing is necessary.

there's a riot at Mustaffa Centre **[Civil Disorder]**. Some bus hit a Bangla **[Traffic]**. Then the other banglas not happy so they start a riot **[Civil Disorder]**

Ambulance & vehicle damage **[Infra/Util Damage]**, 5 CD personnel to the hospital **[Affected Individual]** now and a police car was on fire **[Infra/Util Damage]**!

Figure 1. (a) Crisis-related Tweet Annotated with Situational Details (above); (b) Crisis-related Tweet Annotated with Impact Factors (below)

Actionable Info	Label Type	Description	Label Size[1]
Situational Details	Traffic	vehicle accidents	237
	Fire/Explosion	fire accidents, explosions, wildfires	291
	Flood/Typhoon	floods, heavy rains, typhoons, storms	265
	Civil Disorder	protests, demonstrations, riots, etc.	301
	Armed Assault	shootings, stabbings	261
	Bombing	terrorist bombings	275
Crisis Impact	Affected Individual	casualty, deaths, injuries, etc.	547
	Infrastructure/Utility Damage	property damage, etc.	350
	Infrastructure/Utility Disruption	traffic disruption, power shortage, etc.	81

Table 1. Label Types for Situational Details and Crisis Impact

Two types of actionable information were examined in our study: situational details (SD) and crisis impact (CI). SD such as "what", "who", "when" and "where", was the key enabler of a crisis dispatch while CI estimated crisis severity for ranking and resource allocation. The SD information was categorized into six crisis types (e.g., flood or traffic accident), as each crisis type associated with certain elements and conditions due to specific crisis natures. For example, in terms of location, floods usually hit an entire area while traffic accidents occur at a single spot. It was thus logical to group situational details by crisis types. While there were numerous crisis types, it was infeasible to consider all. After conducting interviews with users and a review of the historical disaster profiles in Singapore (Lai and Tan 2014), six crisis types were identified in the study with an emphasis on civil emergencies such as major fires, flooding caused by torrential rains or typhoons, traffic accidents, bombing explosions, armed assaults, and civil disorders (Table 1). As a result, disasters such as earthquakes, volcanic eruptions, pollutions, or epidemics were not considered. With these six crisis types identified, we annotated spans in crisis-related tweets containing SD such as "what", "who", "when" and "where", which were key components in our definition of actionability. Specifically, we annotated spans that 1) indicated or described the crisis type ("what kind of crisis event occurred?"); 2) mentioned key entities ("who or what was involved in the crisis?"); 3) mentioned crisis location or time ("where and when did it occur?"). Figure 1(a) shows SD annotation on a crisis-related tweet in 2013 Singapore Little India Riot. "Mustaffa Centre" (with a typo) was the location of the "riot". The "riot" and "Mustaffa Centre" were labelled within the same span to ensure the relatedness between the crisis and location, therefore other irrelevant location mentions would be excluded during prediction. A Bangladeshi

[1] Number of annotated spans on crisis tweets used in the study (more in the "Dataset" section).

was the victim of the action "hit" which described the bus accident. In addition, intermediate tokens were included if the resulting span preserved the relation between entities in a sentence. For instance, we labelled "other banglas not happy so they start riot" instead of "other banglas" and "start riot" separately. While most crisis-related tweets belonged to one main crisis type (e.g., "Civil Disorder" in 2013 Singapore Little India Riot), all crisis impact types (Table 1) applied to them regardless of their crisis types. The CI sequence tagging extracted impact factors used to determine crisis severity for prioritization. As various impact classification schemes existed in publications, each with slightly different inclusion and overlapping with one another, we selected CI label types which were both easily quantifiable and relevant to our context. We identified fatality as the most important factor in measuring severity. Other impact factors such as injury and property damage were also extracted, although they were shown to have a strong linear relationship with fatality (Caldera et al. 2016; Caldera and Wirasinghe 2021). Subsequently, we annotated spans in crisis-related tweets that 1) indicated or described any defined impact factor; 2) mentioned key entities (e.g., name of affected individuals or damaged objects); 3) mentioned the crisis severity in numerical values (e.g., '4 injured'). Figure 1(b) shows CI annotation on a crisis-related tweet in 2013 Singapore Little India Riot.

Multi-Task Learning

We proposed a Multi-Task Learning (MTL) approach in which SD and CI sequence tagging tasks were learnt simultaneously via parameter sharing and a joint loss function. MTL has achieved remarkable success in NLP applications, thanks to its improved data efficiency and reduced overfitting (Zhang and Yang 2017). It also improved inference speed as multiple tasks are performed jointly, instead of sequentially. This was a desirable feature especially in the time-critical context of social media crisis informatics.

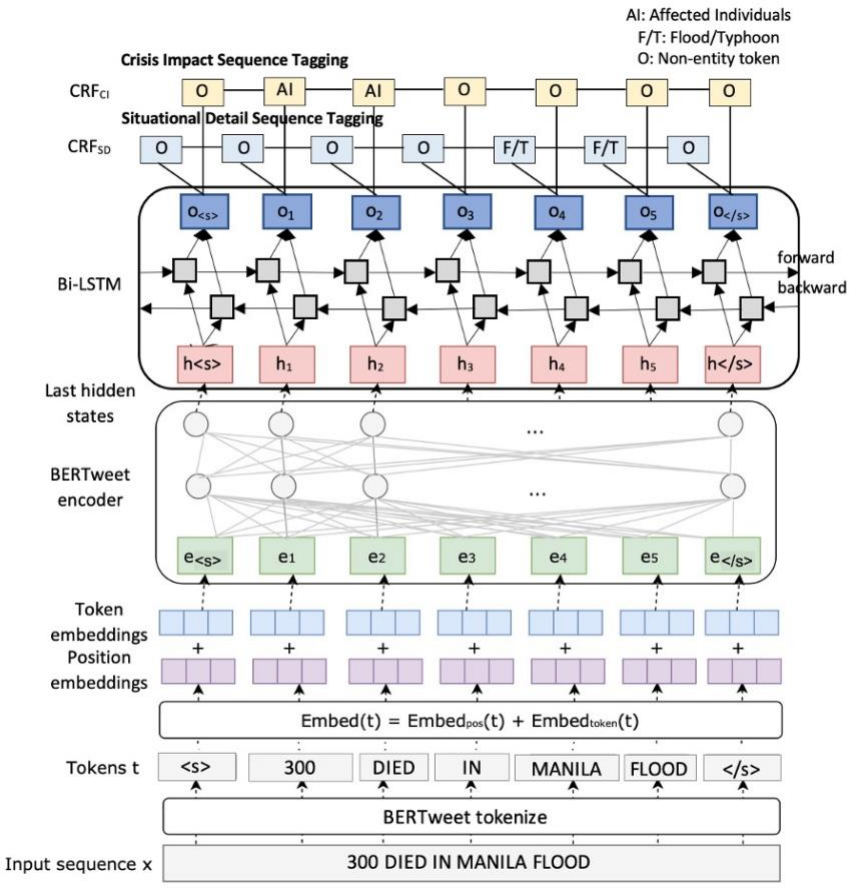

Figure 2. Multi-Task Learning Architecture

Figure 2 depicts the overview architecture of the MTL approach. The architecture comprised of three layers, namely BERTweet embedding layer, bidirectional Long-Short Term Memory (bi-LSTM) layer, and the CRF decoder layer. The BERTweet layer tokenized and generated contextualized vector representation of input sequences. The output of BERTweet was used as the input for the bi-LSTM layer. The bi-LSTM network can efficiently capture past and future features (via forward and backward states) to predict token labels. In this study, bi-LSTM was referred to as LSTM for ease of reference since all LSTMs implemented were bidirectional. Lastly, the CRF decoder extracted sentence level tag information by computing transition compatibility between all pairs of labels on neighboring tokens. The CRF decoders were represented by horizontal lines which connected consecutive output layers. The BERTweet and LSTM layers were shared by two tasks. For illustration purpose, we have used a short tweet "300 DIED IN MANILA FLOOD" in Figure 2. The SD span (light blue) was "MANILA FLOOD" (Flood/Typhoon) and CI span (yellow) was "300 DIED" (Affected Individual).

Experiment

Dataset

Crisis Type (tweet size)	Crisis Event (tweet size)
Traffic (191)	mixed traffic crashes (50), 2013 Glasgow Helicopter Crash (48), 2013 NYC Train Crash (48), 2013 Lac-Megantic Train Crash (45)
Fire/Explosion (189)	2013 West Texas Explosion (43), 2013 Brazil Nightclub Fire (46), 2012 Colorado Wildfires (36), 2019 Durham Gas Explosion (20), 2016 Puttingal Temple Explosion (22), 2017 Lilac Wildfire (22)
Flood/Typhoon (186)	2012 Typhoon Pablo (49), 2013 Alberta Floods (34), 2013 Typhoon Yolanda (38), 2020 Edenville Dam Failure (20), 2013 Queensland Floods (23), 2018 Hurricane Florence (22)
Civil Disorder (191)	mixed civil unrests from 42 countries (87), 2013 Singapore Little India Riot (39), 2020 U.S. Capitol Riot (20), 2022 Iran Protests (45)
Shooting (187)	2013 LA Airport Shootings (34), 2020 South Carolina Bar Shooting (35), 2020 Texas University Shooting (35), 2018 Pittsburgh Synagogue Shooting (39), 2017 Dallas Shooting (44)
Bombing (189)	2016 Brussels Bombings (34), 2017 Manchester Arena Bombing (31), 2013 Boston Bombings (45), mixed bombings on social media (33), 2015 Paris Attacks (46)
Noisy Tweets (1133)	Not crisis-related, randomly sampled (1133)

Table 2. Statistics of Twitter Crisis Dataset in the Study

We collected crisis-related tweets from CrisisLexT26 (Imran et al. 2013), Traffic Tweets (Dabiri 2018), Disasters on Social Media (Crowdflower 2015), TREC-IS (McCreadie et al. 2019; McCreadie et al. 2020), Civil Unrests on Twitter (Sech et al. 2020) and Twitter API (Table 2). A balanced distribution of tweet sizes among crisis types was ensured in our crisis dataset. Within each crisis type, multiple crisis events were included to increase the linguistic variations and train a more generalized model. In total, 1133 crisis-related tweets were collected and annotated with situational details and crisis impact spans using the annotation rules mentioned in the previous section. The label sizes were shown in Table 1 (in "Sequence Tagging and Annotation"). The label size was not to be confused with tweet size as one tweet might contain one or more labelled spans. Table 1 showed a somewhat balanced label size distribution among situational details while for crisis impact it was highly skewed towards "Affected Individual" and "Infrastructure/Utility Damage". Lastly, the dataset was augmented by another 1133 randomly sampled, irrelevant tweets which could be of any topic and contained no actionable information. Given the moderate size of our dataset, the addition of negative samples might improve model generalization. While the impact of negative samples on social

media crisis detection has not been examined sufficiently, random noise was added in the train dataset since it has been shown to regularize overfitting (Zhang et al. 2023).

Model Configuration and Training

BERTweet tokenizer employed a normalization strategy in which Twitter-specific tokens such as user mentions and web links were transformed into special tokens "@USER" and "HTTPURL". Emoticons were converted into corresponding texts. No additional tweet pre-processing was required. This tweet normalization strategy was applied to all models examined in the experiment. Two baseline models (GloVe-LSTM and BERT-FC) were implemented. The first was a LSTM model with a static word embedding GloVe trained on tweets ("glove.twitter.27B.200d"). The second was a BERT ("bert-base-cased") with a fully connected (FC) layer for token classification. For BERTweet models, we implemented both "bertweet-base" and "bertweet-large". The maximum token length was set to 128. All LSTM models had one forward and one backward layer, with a hidden dimension of 128 and a dropout of 0.25. All model configurations were by default unless otherwise stated. CRF was used as the tag decoder for all models.

Given the moderate-sized dataset, a 5-fold cross-validation (CV) was used for model selection. For each CV iteration, the dataset was divided into 80% train and 20% test. During hyperparameter tuning, the 80% train was further divided into 70% train set for training and 10% validation set for evaluation. We used grid search over learning rates {5e-4, 1e-4, 5e-5, 2e-5} and batch sizes {1, 2, 4, 8, 16, 32} for transformer-based models. For the GloVe model, we used grid search over learning rates {0.1, 0.05, 0.01, 0.005, 0.001} and batch sizes {1, 2, 4, 8, 16}. The loss function was defined as the weighted sum of each loss function (cross-entropy) in the MTL, therefore both tasks were learnt simultaneously via backpropagation. The weights were determined by their homoscedastic uncertainty (Cipolla et al. 2018). Model parameters were optimized using AdamW (Loshchilov and Hutter 2017) optimizer and a linear scheduler with warm up steps equal to 1/10 of total training steps. The training would stop if the harmonic mean of $F1_{SD}$ and $F1_{CI}$ scores on validation or test split did not improve for 4 epochs or if the maximum number of 40 epochs was reached. For each CV iteration, once the best hyperparameters were found, the model was trained with both train and validation set before being evaluated on 20% test set. The average Precision, Recall and F1 scores were recorded across the 5-fold cross-validation. They were defined as:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Result

We compared the Precision, Recall and F1 scores for CI and SD sequence tagging between baseline and transformer-LSTM models (Table 3). Between the two baseline models, BERT-FC significantly outperformed GloVe-LSTM. This showed that BERT was superior to LSTM at encoding contextual information, likely due to BERT's self-attention mechanism and the benefit of transfer learning as BERT was pretrained on a large unlabelled corpus. On the other hand, simple LSTM has been shown to outperform BERT on small datasets on which BERT tends to overfit (Kabbara and Cheung 2021; Henlein and Mehler 2022). This likely explained the improvement of BERT-LSTM over BERT-FC when the fully connected layer was replaced by a LSTM encoder. This showed that even in the presence of BERT, using a LSTM encoder might be beneficial. Among transformer-LSTM models, BERTweet-large achieved the best Precision, Recall, and F1 scores. While BERT was pretrained on structured texts such as books and Wikipedia articles, BERTweet was pretrained on tweets therefore showed sizeable improvement over the standard BERT on both Twitter sequence tagging tasks.

While sequence tagging has its advantages, classifying each token is undoubtedly a tougher task than classifying the entire sequence. As a result, sequence tagging usually leads to lower performance than classification on the same dataset (Zhang et al. 2023). It is worth mentioning that the leading F1 score of Twitter NER (a sequence tagging task) is 59.5% on WNUT 2016 NER dataset (Hu et al. 2022). The relatively

low performance (best F1 scores of ~65% for both tasks) suggests the need to include a human-in-the-loop (HITL) mechanism to continuously improve the model through human feedbacks in the future work (Snyder et al. 2019). In addition, by placing equal importance on Precision and Recall, we used the F1 score as the main model selection metric. However, in practice, crisis responders might prefer recall over precision if they primarily care about seeing all valuable information, i.e., missing actionable information is more serious than seeing irrelevant data (McCreadie et al. 2019). Therefore, we recognized that the evaluation metrics for hyperparameter tuning and model selection should be user dependent. Table 4 and 5 showed the BERTweet-large scores by label types in each task. "Infrastructure/Utility Disruption" had the lowest scores among impact types, possibly due to its small label size. The imbalanced CI label type distribution was highly skewed towards "Affected Individual" (see Table 1). While class distribution among SD label types was somewhat balanced, "Civil Disorder" was predicted poorly compared to other classes (Table 5). This was likely due to its broad definition (including protests, riots, demonstrations, etc.) which led to large linguistic variations. We plan to address both issues by increasing the data size via semi-supervised learning in the future (Alam et al. 2018; Sirbu et al. 2022).

Embedder	Encoder	CI			SD		
		Precision	Recall	F1	Precision	Recall	F1
GloVe	LSTM	43.79	29.72	35.35	40.06	27.85	32.69
BERT	FC	56.96	52.85	54.67	58.32	58.50	58.40
BERT	LSTM	58.42	56.16	57.24	59.59	59.96	59.76
BERTweet-base	LSTM	61.19	60.16	60.58	61.57	61.88	61.71
BERTweet-large	LSTM	**63.08**	**65.92**	**64.46**	**65.44**	**65.49**	**65.44**

Table 3. Crisis Impact and Situational Details Sequence Tagging Results

Crisis Impact	Precision	Recall	F1
Affected Individual	64.28	68.09	66.10
Infrastructure/Utility Damage	64.69	66.56	65.50
Infrastructure/Utility Disruption	51.74	50.62	50.74

Table 4. BERTweet-large CI Sequence Tagging Results

Situational Details	Precision	Recall	F1
Bombing	63.83	66.23	64.90
Civil Disorder	46.12	42.91	44.43
Armed Assault	70.93	69.86	70.33
Fire/Explosion	70.57	71.29	70.78
Flood/Typhoon	65.14	66.60	65.81
Traffic	74.53	75.83	74.98

Table 5. BERTweet-large SD Sequence Tagging Results

System Design

We proposed a Crisis Actionability Extraction System (CAES) based on the MTL sequence tagger trained in the previous section. Figure 3 depicts the overall system workflow. The system took tweets collected via Twitter API over a period T as input. The MTL sequence tagger processed and labelled tweets with CI and SD spans. We filtered crisis-related tweets by looking at the presence of any CI or SD span. Tweets without

any tagged span did not contain actionable information and were discarded. After the extraction of actionable phrases via sequence tagging, additional modules, i.e., severity extraction, NER, and localization, were used to measure crisis severity using impact factors and determine the specific aspects of situational details like crisis location. In the following sub-sections, we briefly explained the functionalities of these modules as their implementation was not the focus of this paper. Lastly, crisis-related tweets over the period T were grouped into separate events using crisis type, location, and time as more than one crisis may happen during T. The actionable information related to each event was then aggregated and presented in descending order of severity to crisis responders in tabular form.

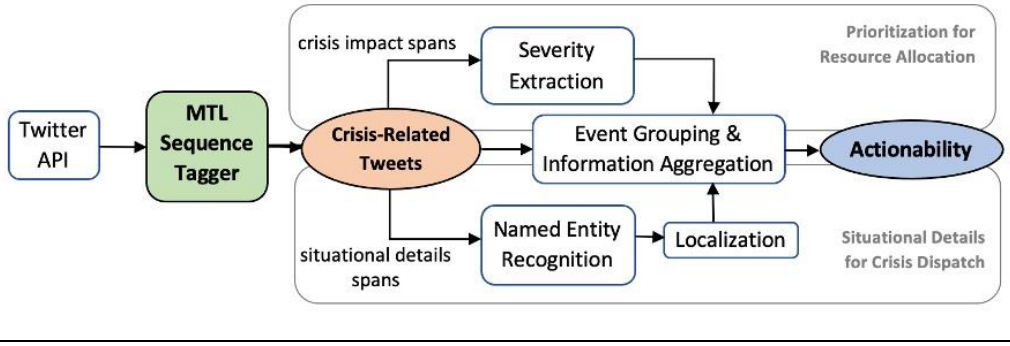

Figure 3. CAES Workflow Diagram

Severity Extraction

The aim of severity extraction is to quantify each impact factor. The main functionality of the module was to extract numerical values associated with keywords in each impact type by pattern matching. For example, keywords associated with fatality may include "death", "died", "killed", and "fatality". Given CI spans such as "3 killed" and "5 ppl died", the fatality count of 3 and 5 can be extracted respectively. These two CI spans may be found in the same tweet, or more likely in two different tweets, with one posted later than the other. To resolve the value inconsistency, the numerical values for fatality, injuries, and monetary damage (if mentioned) from all crisis-related tweet in the same crisis event were later aggregated and the number with the latest timestamp was used for each impact factor. For impact factors that were difficult to quantify, e.g., infrastructure disruptions such as "airport closed" and "road blocked", we proposed to use the mention count as the value. As each impact factor was associated with a numerical value, the overall crisis severity can be calculated by assigning them specific weights (see "Information Aggregation" section).

Named Entity Recognition and Localization

The Named Entity Recognition (NER) module was used to extract named entities such as PER (person), ORG (organization), LOC (location), and FAC (building name) within SD spans. For reasonable accuracy on tweets, we used an off-the-shelf BERTweet-based NER model trained using Tweebank-NER (Jiang et al. 2022), which was a training corpus for Twitter NER. Once location-related named entities, specifically LOC and FAC, were extracted via NER from SD spans, normalization to standard forms was required as abbreviations and typos were common on Twitter. For example, "Ang Mo Kio", a district in Singapore, was often abbreviated as "AMK". A list of place abbreviations in Singapore was collected and a geolocation-based Knowledge Base (KB) was constructed with three levels of granularity: area, street, and building (Figure 4). The normalized location was matched to an entry in the KB via string matching. For example, the abbreviated term "AMK St 53" was normalized to the standard form "Ang Mo Kio Street 53" and mapped to the entry at a street level. The geolocation granularity enabled information aggregation of crisis-related tweets within a specific area, street or building, depending on the use case. The practitioners may build a similar geolocation-based KB according to the proposed structure in Figure 4.

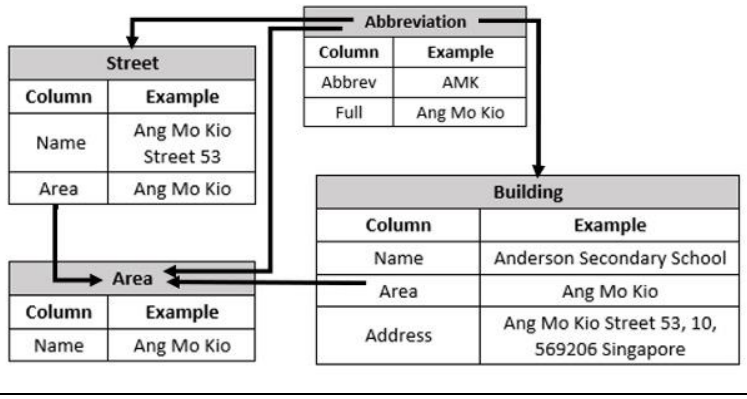

Figure 4. Proposed Structure for Geolocation Knowledge Base

Event Grouping and Information Aggregation

After the extraction of situational details and crisis impact, crisis-related tweets were grouped into crisis events by crisis type, location, and time (tweet timestamp). For instance, in the detection period T, all crisis-related tweets of "Fire/Explosion" (F/E) at location A were grouped under the same event "F/E at A during T". The detection period T can be set according to the use case (e.g., 30 minutes). The location A, as mentioned before, can be set to any aggregation level among "area", "street", or "building". If it was set to "area", for example, crisis-related tweets of F/E in Ang Mo Kio were grouped under the event named "F/E at Ang Mo Kio during T". As some crisis-related tweets might contain SD spans of multiple crisis types, we allowed such tweets to be included in multiple events at the same time. Upon event grouping, impact values and situational details were aggregated from all crisis-related tweets belonging to a crisis event. Note that the severity could change as the crisis developed. Even in the same detection period, different fatality counts could be extracted from tweets posted at different times. In the case of different values, we used the ones extracted at a later timestamp. Once the numerical values for impact factors were aggregated from the tweets (e.g., 5 fatalities, 20 injuries), we defined a crisis severity score by computing the weighted sum of these impact factors:

$$\text{Severity} = w_1 * \text{Fatalities} + w_2 * \text{Injuries} + w_3 * \text{Damage Cost} + w_4 * \text{No. of Entity Mentions} + w_5 * \text{No. of Tweets}$$

These weights should be customized by crisis practitioners based on their use cases. In addition to impact factors, we also included the number of crisis-related tweets as a factor in severity measurement as tweet count usually suggested the significance of an event. Finally, crises were ranked according to severity and presented to the user in tabular form with situational details. In the process of information aggregation, there could be instances where certain situational details (e.g., location) might vary or contradict each other. When dealing with different locations extracted from the same event, we resolved the inconsistency by employing a majority voting approach in which the location with the most occurrences or "votes" was used as the final, agreed-upon crisis location.

Case Study: 2013 Singapore Little India Riot

The 2013 Singapore Little India Riot was the second riot in post-independence Singapore. It took place on 8 December 2013, at 21:23 local time, after a private bus ran over and killed an Indian construction worker at the junction of Race Course Road and Hampshire Road in Little India, Singapore. Soon after the accident, the crowd started to form at the scene. Police reinforcements arrived progressively at around 21:45. The rioters attacked the private bus and emergency vehicles. The Special Operations Command (SOC) was activated and arrived at about 22:30. At 22:44, the SOC forces began to disperse the mob and arrest rioters. About 300 migrant labourers were involved in the riot. Twenty-five emergency vehicles were damaged in the riots, alongside five that were set on fire. Since there were hundreds of thousands of tweets mentioning this crisis, we showed some sampled tweets labelled by the MTL sequence tagger in Table 6. The earliest detected tweet was at 21:38. It reported a "Civil Disorder" crisis, specifically, a riot. Situational details

extracted at 21:52 provided key entities of the event, i.e., "police", "ambulance", "crowd", and most importantly, the location "#littleindia". More location mentions followed: "Little India" and "Race Course Road" at 22:42 and "Mustaffa Centre" (with a typo) at 22:51 (Mustafa Centre was a mall located in Little India, Singapore). The CI spans showed up later than SD spans. A range of impact factors such as road blockage, vehicle damage, fatality and injuries were reported.

Time	Tweet	SD spans	CI spans
21:38	Lol after star parade straight riot sad	"riot" [Civil Disorder]	-
21:52	Police, ambulance, crowd screaming again at #littleIndia	"Police, ambulance, crowd screaming again at #littleIndia" [Civil Disorder]	-
22:35	More police cars, louder screaming, some kind of gun firing, road blocked…chaos	"More police cars", "gun firing" [Civil Disorder]	"road blocked" [Infra/Util Disruption]
22:42	There's a riot at Little India- police cars set on fire and turned up side down? Call in the Gurkhas.	"a riot at Little India" [Civil Disorder]	"police cars set on fire and turned up side down" [Infra/Util Damage]
22:42	Riot at Race Course Road HTTPURL	"Riot at Race Course Road" [Civil Disorder]	-
22:51	there's a riot at Mustaffa Centre. Ambulance & vehicle cd damage, 5 CD personnel to the hospital now and a police car was on fire!	"a riot at Mustaffa Centre" [Civil Disorder]	"Ambulance & vehicle cd damage", "a police car was on fire" [Infra/Util Damage]; "5 CD personnel to the hospital" [Affected Individual]
22:59	My dad just called - 1 guy died, 3 police cars on fire, ambulances on fire, entire Race Course road is chaotic. People are rioting	"People are rioting" [Civil Disorder]	"1 guy died" [Affected Individual]; "3 police cars on fire", "ambulances on fire" [Infra/Util Damage]; "entire Race Course road" [Infra/Util Disruption]

Table 6. Sampled Tweets with Labelled SD and CI Spans

The location mentions extracted via NER were normalized and matched them to an entry in the KB. For example, "#littleindia" or "Little India" was a subzone in the area of Rochor. "Mustaffa Centre" (with a typo) was matched to "Mustafa Centre" which was a building entry in the KB. "Race Course Road" was the street in Rochor where the riot first took place. We set the location level to "area". For a given time interval, e.g., 23:00 to 23:30, we grouped crisis-related tweets in the area of Rochor by their crisis types. Table 7 showed the event groupings alongside with actionable information. The situational details and impact values were aggregated after removing duplicates. Note that tweets were actually grouped into three separate events instead of one. This was because some tweets were labelled with "Traffic" or "Fire/Explosion" spans, reflecting the complexity of the underlying crisis, which was a riot caused by a fatal traffic accident where angry mobs set vehicles on fire. The main crisis event "Civil Disorder at Rochor 23:00 to 23:30" consisted of 133 crisis-related tweets with actionable information out of hundreds of thousands of tweets. Since the severity was highest (2.133)[2], it was displayed on top with the highest priority.

[2] For demonstration purpose, the severity scores were calculated using the simplified equation: *Severity = Fatality + 0.1 * Injuries + 10^{-3} *No. of Tweets*

Crisis responders could leverage actionable information in Table 7 to effectively address the crisis situation in a comprehensive manner. By thoroughly assessing the situational details of each incident, including factors such as the cause of the crisis (i.e., bus accident), the specific locations (i.e., Race Course Road), and the impact severity (e.g., fatalities and injuries), responders could gain a clear understanding of the ongoing crisis scenario. This understanding was crucial for the strategic resource allocation and crisis mitigation. By looking at the riot impact (e.g., burning of emergency vehicles), responders could prioritize the deployment of back-up emergency vehicles and reinforcement to address the immediate danger and prevent further escalation. The crisis impact, "10 police officers were injured", could serve as an indicator for responders' safety, warranting preparatory measures like protective gear when handling the riot situation. With insights into disruptions in transportation and traffic flow resulting from the riot, responders could proactively manage traffic routes and reroute bus services in the affected areas. Effective collaboration and communication among different response units, including law enforcement, ambulance services, and fire departments, could be facilitated through the shared details of the riot, enabling a coordinated and efficient approach to crisis management. Public safety could also be enhanced, as responders can disseminate information about the ongoing crisis, locations affected, and potential risks to the public, thereby minimizing exposure to danger. Overall, by utilizing the extracted actionable information, crisis responders could make well-informed decisions, mitigate risks, and effectively navigate the complexities of the crisis, ultimately ensuring the safety and well-being of both responders and the community at large.

Event	No. of Tweets	Actionable Information
Civil Disorder at Rochor	133	Situational details: A riot started when a bus driver hit and killed an Indian national at (Race Course Road, Mustafa Centre) in Little India, Rocher. Police in riot gear. Mob gathered. At least 200 people are rioting. Impact: Emergency vehicles including police cars and ambulance were damaged. At least one police car was overturned and ambulance set on fire by rioters. SOC (riot police) was activated. 10 police officers were injured. Bus services passing by or through Little India were stopped. Severity: 1 fatality. 10 injuries. (score: 2.133)
Traffic at Rocher	6	Situational details: A bus driver knocked down and killed an Indian man at Race Course Road, Rocher Impact: A riot started in Rocher when a bus driver hit someone. Severity: 1 fatality (score: 1.006)
Fire/Explosion at Rocher	8	Situational details: rioters set fire on emergency vehicles including ambulances; vehicles on fire; bus burning in Little India Impact: vehicle damage, riot gear deployed Severity: - (score: 0.008)

Table 7. Crisis Events between 23:00 and 23:30

Challenges and Future Work

In our paper, we grouped tweets by crisis type, location, and time. While clustering is another common method used to aggregate tweets, it brings complexity to the real-world application as the word frequencies change dramatically over time on social media. The short, noisy nature of tweets also makes representation more difficult, thus increasing the clustering error. The unknown number and density of event clusters add to the challenge. In our early exploratory experiment, HDBSCAN (Hierarchical Clustering and Density-Based Spatial Clustering of Applications with Noise) was used as the clustering technique to group similar tweets (Campello et al. 2013). It was chosen as it did not require the number of clusters in advance and produced a clustering that gave the best stability. However, we found that HDBSCAN was unable to separate tweets of the same crisis type but belonging to different events. For example, the following two "Traffic" tweets were assigned to the same cluster when they mentioned two distinct events: "3 dead following early morning crash along AYE" and "1 dead, 1 injured in motorcycle crash in Bt Batok". The advantage of our

event grouping technique was that it could separate similar events that happened during the same interval. In addition, it enabled users to define the location level of grouping (area, street and building) providing better flexibility in real-world application. We found that using the "what", "when" and "where" aspects as the grouping criteria was both simple and effective, given that our system could accurately extract such information. The downside of this method was that relevant tweets without location information were left out from the event groups. They might contain additional information that would be lost in the Information Aggregation. If users wish to capture these "left-out" tweets, an additional component could be implemented by using the semantic search, that is, comparing the semantic similarity between the average event grouping embedding and each of these "left-out" tweets. If the similarity exceeds a certain threshold, the tweet can be added to the group. We have showed in Table 7 that CAES was capable of extracting multiple crisis types, for example 'Civil Disorder', 'Traffic', and 'Fire/Explosion' from the Little India Riot crisis. Even though we grouped crisis-related tweets based on the "what" (crisis type), "when" (detection period), and "where" (localization with KB) attributes of an event, it could be challenging to determine if separate event groups were referring to different events or different aspects of the same event. We recognize the potential limitation and suggest that it can be beneficial to include a human-in-the-loop verification step in the future. Information reliability is another major problem. Ensuring the information reliability entails fact-checking which includes removing rumors and false information, aggregating authentic information from various sources and cross-checking details. This process is both time-critical and labor-intensive. For future work, we plan to utilize the capabilities of a ready-to-use, general Question Answering (QA) model such as Macaw (Multi-angle c(q)uestion answering) (Tafjord and Clark 2021) to handle the task of fact-checking. Lastly, due to the lack of a universal crisis severity classification scheme, it was challenging to standardize severity measurement. Although we attempted to include impact factors such as infrastructure/utility damage and disruption, e.g., "car overturned" and "road blocked", their numerical values were either rarely reported or hard to quantify. Among impact factors, the fatality was used as a main factor to measure crisis severity. The good news was that a strong linear relationship was shown to exist between fatality and other impact factors. For instance, an increase in the number of fatalities predicted an increased in number of injuries and damage (Caldera et al. 2016; Caldera and Wirasinghe 2021). This might suggest that fatality alone is sufficient to measure severity. Nonetheless, more research needs to be done and crisis practitioners should determine the severity measurement based on their use cases.

Conclusion

In this paper, we aimed to bridge the gap in existing crisis management tools on social media by implementing actionability in the system design. Compared to most systems that only identified actionable messages, we proposed a Crisis Actionability Extraction System (CAES) which was the first social media processing system that handled filtering, classification, phrase extraction, severity estimation, localization, and aggregation of actionable information altogether. In the process of system development, we examined the effectiveness of a BERTweet-LSTM-CRF model via Multi-Task Learning to simultaneously extract two types of actionable information: situational details and crisis impact. The extracted actionable information was then processed in CAES and presented to crisis responders. We demonstrated workflow of CAES in actionability extraction using a case study of 2013 Singapore Little India Riot and highlighted its usefulness in reducing cognitive load of crisis responders, mitigating risks, and navigating the complexities of the crisis. We believe that the proposed system design and models can be adapted for other social media platforms and the implementation is not limited to Twitter.

Acknowledgements

The Authors would like to acknowledge the support and project funding from ST Engineering Mission Software & Services Pte Ltd under Research Collaboration Agreement No: 001052-00001.

References

Alam, F., Joty, S., and Imran, M. 2018. "Graph Based Semi-supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets," in Proceedings of the International AAAI Conference on Web and Social Media (12:1).

Alam, F., Sajjad, H., Imran, M., and Ofli, F. 2021. "CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing," in Proceedings of the International AAAI Conference on Web and Social Media (15:1), pp. 923–932.

Al-Zaidy, R. A., Caragea, C., and Giles, C. L. 2019. "Bi-LSTM-CRF Sequence Labeling for Keyphrase Extraction from Scholarly Documents," The World Wide Web Conference.

Ampili, K., and Kanakala, S. 2022. "Tweet Summarization Using Clustering Mechanisms," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 1481-1486.

Caldera, H.J., Wirasinghe, S.C., and Zanzotto, L. 2016. "An Approach to Classification of Natural Disasters by Severity," in Proceedings of the 5th International Natural Disaster Mitigation Specialty Conference, Annual Conference of the Canadian Society for Civil Engineering, London, Canada, NDM-528, pp. 1–11.

Caldera, H. J., and Wirasinghe, S. C. 2021. "A Universal Severity Classification for Natural Disasters," Natural Hazards (111:2), pp. 1533–1573.

Campello, R. J., Moulavi, D., and Sander, J. 2013. "Density-based Clustering Based on Hierarchical Density Estimates," Advances in Knowledge Discovery and Data Mining, pp. 160–172.

Chen, Z., Sun, H., Zhang, W., Xu, C., Mao, Q., and Chen, P. 2023. "Neural-Hidden-CRF: A Robust Weakly-Supervised Sequence Labeler," in Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

Cipolla, R., Gal, Y., and Kendall, A. 2018. "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Coche, J., Montarnal, A., Kropczynski, J., Tapia, A., and Benaben, F. 2021. "Actionability in a Situational Awareness World: Implications for Social Media Processing System Design," in Proceedings of the 18th International Conference on Information Systems for Crisis Response and Management, Vol. 2021-May, pp. 994-1001.

Dabiri, S. 2018. "Tweets with Traffic-related Labels for Developing a Twitter-based Traffic Information System," Mendeley Data, March 23. (https://data.mendeley.com/datasets/c3xvj5snvv/1; accessed May 3, 2023).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.

"Disasters on social media - dataset by Crowdflower." 2016. data.world, November 21. (https://data.world/crowdflower/disasters-on-social-media; accessed May 3, 2023).

Domala, J., Dogra, M., Masrani, V., Fernandes, D., D'souza, K., Fernandes, D., and Carvalho, T. 2020. "Automated Identification of Disaster News for Crisis Management Using Machine Learning and Natural Language Processing," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC).

Fan, Z., Wu, Z., Dai, X., Huang, S., and Chen, J. 2019. "Target-oriented Opinion Words Extraction with Target-fused Neural Sequence Labeling," North American Chapter of the Association for Computational Linguistics.

Garg, P. K., Chakraborty, R., and Dandapat, S. K. 2023. "OntoDSumm: Ontology-Based Tweet Summarization for Disaster Events," IEEE Transactions on Computational Social Systems. pp. 1-16.

Grace, R. 2021. "Toponym Usage in Social Media in Emergencies," International Journal of Disaster Risk Reduction (52), p. 101923.

Graves, A., Mohamed, A.-R., and Hinton, G. 2013. "Speech Recognition with Deep Recurrent Neural Networks," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.

Gidwani, M., and Rao, A. 2023. "Comparative Analysis of Rumour Detection on Social Media Using Different Classifiers.," Informatics and Automation (22:4), pp. 777-794.

He, Z., Wang, Z., Wei, W., Feng, S., Mao, X., and Jiang, S. 2020. "A Survey on Recent Advances in Sequence Labeling from Deep Learning Models," arXiv.org, November 13. (https://arxiv.org/abs/2011.06727; accessed May 4, 2023).

Henlein, A., and Mehler, A. "What do Toothbrushes do in the Kitchen? How Transformers Think our World is Structured," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5791-5807.

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V., and García Villalba, L. 2019. "Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation," Sensors (19:7), MDPI AG, p. 1746.

Hiltz, S. R., Hughes, A. L., Imran, M., Plotnick, L., Power, R., and Turoff, M. 2020. "Exploring the Usefulness and Feasibility of Software Requirements for Social Media Use in Emergency Management," International Journal of Disaster Risk Reduction (42), p. 101367.

Hu, J., Shen, Y., Liu, Y., Wan, X., and Chang, T.-H. 2022. "Hero-Gang Neural Model for Named Entity Recognition," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1924-1936.

Hu, Y., and Wang J. 2020. "How Do People Describe Locations During a Natural Disaster: An Analysis of Tweets from Hurricane Harvey", Leibniz International Proceedings in Informatics, p. 177.

Jiang, H., Hua, Y., Beeferman, D., and Roy, D. 2022. "Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis," in Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pp. 7199–7208.

Kabbara, J., and Cheung, J. C. K. 2021. "Post-Editing Extractive Summaries by Definiteness Prediction," in Findings of the Association for Computational Linguistics (EMNLP 2021), pp. 3682–3692.

Krishnan, H., Roy, A., Menon, A. K., D. S and Babu, H. M. 2023. "Natural Disaster Detection Using Social Media," 2023 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), pp. 1-6.

Kropczynski, J., Grace, R., Coche, J., Halse, S., Obeysekare, E., Montarnal, A., Benaben, F., and Tapia, A. 2018. "Identifying Actionable Information on Social Media for Emergency Dispatch," in Proceedings of the 1st International Conference on Information Systems for Crisis Response and Management Asia Pacific, Wellington, New Zealand, p.428-438.

Kruspe, A., Kersten, J., and Klan, F. 2021. "Review Article: Detection of Actionable Tweets In Crisis Events," Natural Hazards and Earth System Sciences (21:6), pp. 1825–1845.

Lafferty, J., McCallum, A., and Pereira, F. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01), San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, pp. 282–289.

Lai, A. Y.-H., and Tan, S. L. 2014. "Impact of Disasters and Disaster Risk Management in Singapore: A Case Study of Singapore's Experience in Fighting the SARS Epidemic," Resilience and Recovery in Asian Disasters, pp. 309–336.

Li, J., Sun, A., Han, J., and Li, C. 2023. "A Survey on Deep Learning for Named Entity Recognition: Extended Abstract," 2023 IEEE 39th International Conference on Data Engineering, pp. 3817-3818.

Liu, J., Singhal, T., Blessing, L. T. M., Wood, K. L., and Lim, K. H. 2021. "CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding," in Proceedings of the 32nd ACM Conference on Hypertext and Social Media, pp. 133-141.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv.org. (https://arxiv.org/abs/1907.11692; accessed May 4, 2023).

Lo, S. L., Lee, K., and Zhang, Y. 2023. "Is a Pretrained Model the Answer to Situational Awareness Detection on Social Media?" in Proceedings of 56th Hawaii International Conference on System Sciences, pp. 2110-2119.

Loshchilov, I., and Hutter, F. 2019. "Decoupled Weight Decay Regularization," 7th International Conference on Learning Representations.

Luo, Y., Xiao, F., and Hai, Z. 2020. "Hierarchical Contextualized Representation for Named Entity Recognition," 34th AAAI Conference on Artificial Intelligence (AAAI 2020), pp. 8441-8448.

Mazloom, R., Li, H., Caragea, D., Caragea, C., and Imran, M. 2019. "A Hybrid Domain Adaptation Approach for Identifying Crisis-Relevant Tweets," International Journal of Information Systems for Crisis Response and Management (IJISCRAM) (11:2), pp. 1-19.

McCreadie, R., Buntain, C., and Soboroff, I. 2019. "TREC Incident Streams: Finding Actionable Information on Social Media," in Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, Valencia, Spain.

McCreadie, R., Buntain, C., and Soboroff, I. 2020. "Incident Streams 2019: Actionable Insights and How to Find Them," in Proceedings of 17th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2020), pp. 744–760.

Mikolov, T., Chen, K., Corrado, G.S., and Dean, J. 2013. "Efficient Estimation of Word Representations in Vector Space," International Conference on Learning Representations.

Muhammad, S. H., Adelani, D. I., Ahmad, I. S., Abdulmumin, I., Bello, B. S., Choudhury, M., Emezue, C. C., Aremu, A., Abdul, S., and Brazdil, P. 2022. "NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis," International Conference on Language Resources and Evaluation.

Nguyen, D. Q., Vu, T., and Tuan Nguyen, A. 2020. "BERTweet: A pre-trained language model for English Tweets," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 9-14.

Ning, X., Yao, L., Benatallah, B., Zhang, Y., Sheng, Q. Z., and Kanhere, S. S. 2019. "Source-Aware Crisis-Relevant Tweet Identification and Key Information Summarization," ACM Transactions on Internet Technology (19:3), pp. 1–20.

Poblete, B., Guzman, J., Maldonado, J., and Tobar, F. 2018. "Robust Detection of Extreme Events Using Twitter: Worldwide Earthquake Monitoring," IEEE Transactions on Multimedia (20:10), pp. 2551–2561.

Pennington, J., Socher, R., and Manning, C. 2014. "GloVe: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Purohit, H., Castillo, C., Imran, M., and Pandey, R. 2018. "Social-EOC: Serviceability Model to Rank Social Media Requests for Emergency Operation Centers," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 119-126.

Raza, S., and Ding, C. 2022. "Fake news detection based on news content and social contexts: a transformer-based approach," International Journal of Data Science and Analytics (13), pp. 335 – 362.

Reuter, C., Hughes, A. L., and Kaufhold, M.-A. 2018. "Social Media in Crisis Management: An evaluation and analysis of crisis informatics research," International Journal of Human–Computer Interaction (34:4), pp. 280–294 (doi: 10.1080/10447318.2018.1427832).

Rodrigues, A. P., Fernandes, R., A, A., B, A. L., Shetty, A., K, A., Lakshmanna, K., and Shafi, R.M. 2022. "Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques," Computational Intelligence and Neuroscience.

Rudra, K., Goyal, P., Ganguly, N., Mitra, P., and Imran, M. 2018. "Identifying sub-events and summarizing disaster-related information from microblogs," The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. pp.265-274.

Sech, J., DeLucia, A., Buczak, A. L., and Dredze, M. 2020. "Civil unrest on Twitter (cut): A dataset of tweets to support research on civil unrest," in Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020) (doi: 10.18653/v1/2020.wnut-1.28).

Sirbu, I., Sosea, T., Caragea, C., Caragea, D., and Rebedea, T. 2022. "Multimodal Semi-supervised Learning for Disaster Tweet Classification," International Conference on Computational Linguistics.

Snyder, L. S., Lin, Y.-S., Karimzadeh, M., Goldwasser, D., and Ebert, D. S. 2019. "Interactive Learning for Identifying Relevant Tweets to Support Real-time Situational Awareness," in IEEE Transactions on Visualization and Computer Graphics (26:1), pp. 558-568.

Suwaileh, R., Elsayed, T., Imran, M., and Sajjad, H. 2022. "When a disaster happens, we are ready: Location mention recognition from crisis tweets," International Journal of Disaster Risk Reduction (78), p. 103107.

Tafjord, O., and Clark, P. 2021. "General-Purpose Question-Answering with Macaw," arXiv.org, September 6. (https://arxiv.org/abs/2109.02593; accessed May 3, 2023).

Wang, C., Nulty, P., and Lillis, D. 2021. "Transformer-based Multi-task Learning for Disaster Tweet Categorisation," 18th International Conference on Information Systems for Crisis Response and Management.

Xia, C., Zhang, C., Yang, T., Li, Y., Du, N., Wu, X., Fan, W., Ma, F., and Yu, P. S. 2019. "Multi-grained Named Entity Recognition," Annual Meeting of the Association for Computational Linguistics.

Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M., and Starbird, K. 2018. "From Situational Awareness to Actionability: Towards Improving the Utility of Social Media Data for Crisis Response," in Proceedings of the ACM on Human-Computer Interaction (2:CSCW), pp. 1–18.

Zahra, K., Imran, M., and Ostermann, F. O. 2020. "Automatic identification of eyewitness messages on Twitter during disasters," Information Processing & Management (57:1), p. 102107.

Zhang, Y., Lo, S. L., and Myint, P. Y. 2023. "Impact of Difficult Noise on Twitter Crisis Detection," in Proceedings of Pacific Asia Conference on Information Systems (PACIS 2023).

Zhang, Y., and Yang, Q. 2017. "An overview of multi-task learning," National Science Review (5:1), pp. 30–43.