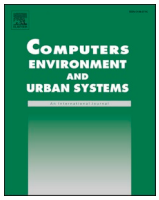
Computers, Environment and Urban Systems 95 (2022) 101824

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



VictimFinder: Harvesting rescue requests in disaster response from social media with BERT

Bing Zhou [a], Lei Zou [a,*], Ali Mostafavi [b], Binbin Lin [a], Mingzheng Yang [a], Nasir Gharaibeh [b], Heng Cai [a], Joynal Abedin [a], Debayan Mandal [a]

[a] Department of Geography, Texas A&M University, College Station, TX, United States of America
[b] Zachry Department of Civil & Environmental Engineering, Texas A&M University, College Station, TX, United States of America

ARTICLE INFO

Keywords:
Social media
Natural language processing
BERT
Disaster response
Emergency rescue

ABSTRACT

Social media platforms are playing increasingly critical roles in disaster response and rescue operations. During emergencies, users can post rescue requests along with their addresses on social media, while volunteers can search for those messages and send help. However, efficiently leveraging social media in rescue operations remains challenging because of the lack of tools to identify rescue request messages on social media automatically and rapidly. Analyzing social media data, such as Twitter data, relies heavily on Natural Language Processing (NLP) algorithms to extract information from texts. The introduction of bidirectional transformers models, such as the Bidirectional Encoder Representations from Transformers (BERT) model, has significantly outperformed previous NLP models in numerous text analysis tasks, providing new opportunities to precisely understand and classify social media data for diverse applications. This study developed and compared ten *VictimFinder* models for identifying rescue request tweets, three based on milestone NLP algorithms and seven BERT-based. A total of 3191 manually labeled disaster-related tweets posted during 2017 Hurricane Harvey were used as the training and testing datasets. We evaluated the performance of each model by classification accuracy, computation cost, and model stability. Experiment results show that all BERT-based models have significantly increased the accuracy of categorizing rescue-related tweets. The best model for identifying rescue request tweets is a customized BERT-based model with a Convolutional Neural Network (CNN) classifier. Its F1-score is 0.919, which outperforms the baseline model by 10.6%. The developed models can promote social media use for rescue operations in future disaster events.

1. Introduction

Natural hazards like hurricanes, tornadoes, and floods are becoming more devastating and increasingly frequent due to climate change (Kryvasheyeu et al., 2016). The emergence of novel forms of big data brings new approaches to understand and mitigate natural disasters' impacts on human communities (Liu et al., 2015). Specifically, the rise of massive social media data enables researchers to view human responses to disasters in near real-time through a special lens. During disasters, social media data reflect how users perceive risks and access information, shaping how they prepare for and respond to hazardous events. Meanwhile, natural hazards spring activities such as pre-disaster evacuation, in-disaster rescue, and post-disaster rebuilding, which could be monitored through social media streams. As a result, the popularity of incorporating social media data and platforms into disaster management continues to grow (Zou, Lam, Cai, and Qiang, 2018; Kirilenko and Stepchenkova, 2014; Zhang et al., 2018; Arthur, Boulton, Shotton and Williams, 2018; Avvenuti, Cresci, La Polla, Marchetti, and Tesconi, 2014).

During Hurricane Harvey in 2017, the local police and fire departments saved only 30% of Houston residents who failed to evacuate and demanded rescue (Gallagher, 2017). When the 911 system was overloaded, many Harvey victims turned to social media for assistance (Mihunov, Lam, Zou, Wang, and Wang, 2020). Houston residents posted rescue requests and their addresses on social media while volunteers searched for those messages and sent help, marking Harvey as one of the first events in which social media played significant roles in fast-response and rescue missions (Rhodan, 2017).

Collecting rescue requests from social media takes three steps (Fig. 1). The first step, referred to as *VictimFinder* in this research, is

* Corresponding author.
E-mail address: lzou@tamu.edu (L. Zou).

https://doi.org/10.1016/j.compenvurbsys.2022.101824
Received 11 November 2021; Received in revised form 4 March 2022; Accepted 10 May 2022
Available online 17 May 2022

identifying rescue request messages from social media data. The second step is toponym recognition, which extracts complete addresses from the detected rescue request messages. Finally, the extracted addresses can be converted to geographical coordinates through geocoding, namely toponym resolution.

However, challenges exist in collecting rescue requests from social media, which hinder the effective application of social media in disaster response and rescue operations. During Hurricane Harvey, social media users lacked knowledge on how to get help online, so they composed rescue request messages differently (Mihunov et al., 2020). Consequently, searching rescue requests in the first step was mainly accomplished manually, requiring intensive human resources and time. There is a need to develop algorithms and tools to automate harvesting rescue requests from social media applications.

The breakthroughs in Natural Language Processing (NLP) provide solutions to the challenges. One way to identify rescue request messages from social media data is through text understanding and classification, which involves using NLP techniques to understand and categorize social media messages. Recently, the introduction of bidirectional transformers models in feature extraction, such as the Bidirectional Encoder Representations from Transformers (BERT) model, has significantly outperformed previous NLP models in numerous text analysis tasks (Devlin, Chang, Lee, and Toutanova, 2019), offering a great potential to develop robust classifiers that can precisely recognize rescue requests from social media data during disasters.

In light of this concept, we developed *VictimFinder* models based on advanced NLP algorithms, including BERT, for recognizing rescue request Twitter messages, referred to as tweets. The primary objective is to examine if BERT-based models can significantly increase the efficacy of identifying rescue request tweets with limited training data. Rescue-related tweets during 2017 Hurricane Harvey were collected and used to train ten *VictimFinder* models, including three based on milestone NLP algorithms, four BERT-based, and three customized BERT-based models. We evaluated each model by classification performance, computation cost, and model stability. The developed optimal models can be applied to harvest rescue requests from social media in future hazard events. Findings from this study will shed light on the potentials and limitations of different language modeling algorithms in analyzing social media data.

The article proceeds as follows. We first provide a brief review of previous investigations on the analysis of social media data for disaster research, a summary of NLP for tweet classification, and an introduction to Hurricane Harvey in Section 2. Section 3 details the methodology of building tweet classification models using different NLP algorithms. The training Twitter data collection and preprocessing, experiment implementation, and model evaluation are explained in Section 4. Following that, we document the results in Section 5. Finally, we conclude with a summary of the findings and their implications in Section 6 while discussing the methodological uncertainties and limitations of the study and providing suggestions for future research.

2. Background

2.1. Social media analysis for disaster research

The emergence and rapid development of information and communications technology (ICT) have turned individuals into sensors, fostering the production of human-generated geospatial big data. Such datasets bring new channels for us to gain deeper understandings of the socioeconomic environment at multiple spatial and temporal scales. The use of geospatial big data to study socioeconomic characteristics is defined as social sensing (Liu et al., 2015). Taxi trajectories, mobile phone records, location data recorded by mobile sensors, social media, and social networking data are popular forms of geospatial big data (Liu et al., 2015).

In terms of disasters, previous research has been dependent on traditional data collected at regular intervals, for example, data from surveys, health agencies, and the census. These data are usually collected by reliable authorities or research teams. Mihunov et al. (2018) developed a resilience inference measurement (RIM) framework to estimate county-level resilience scores to drought by leveraging socioeconomic data from the U.S. Census and health data from the U.S. Department of Health and Human Services. Socioeconomic data can also be incorporated with inundated areas detected from remote sensing to assess disaster damages (Qi and Altinakar, 2011). Online or telephone surveys are commonly used to collect information on post-disaster societal impacts, such as investigating the effect of flood risk on migration considerations in coastal Louisiana (Correll, Lam, Mihunov, Zou, and Cai, 2021). However, these data are unable to describe communities' real-time preparedness, response, and recovery behaviors during hazardous events (Zou et al., 2018). Such drawbacks can be tackled by taking advantage of social sensing. Social media, for example, have gradually been integrated into our daily lives and become a major contributor of effective social sensing, offering significant potentials in disaster management.

With the fact-finding accuracy of analysis based on social media continues to progress (Wang et al., 2015), and social media data resources have burgeoned, they have been applied to investigate different types of natural hazards. For example, Avvenuti et al. (2014) implemented an early earthquake detecting and warning system using Twitter data, which offers timely detection of events in Italy with a False Positive rate of 10% regarding earthquake magnitude over 3.5 Richter. A National Landslide Database sourced from social media since 2012 has been constructed and publicized that underpins future landslide forecast and assessment (Pennington, Socher, and Manning, 2014). Wildfire-related tweets have been analyzed to reveal the situational and geographical awareness of the users (Wang, Ye, and Tsou, 2016). Agile monitoring of when and where the flood takes place is achieved by combining remote sensing data and social media (Jongman, Wagemaker, Romero, and De Perez, 2015). Monitoring social media data also helps detect disaster event centers rapidly so that responding agencies can task satellite observations on those areas for more accurate disaster understanding and response (Cervone et al., 2016). Several case studies processed social media data by geocoding and sentiment analysis tools to analyze the spatial patterns of changing public awareness and emotions toward hurricanes in different phases of the disaster management

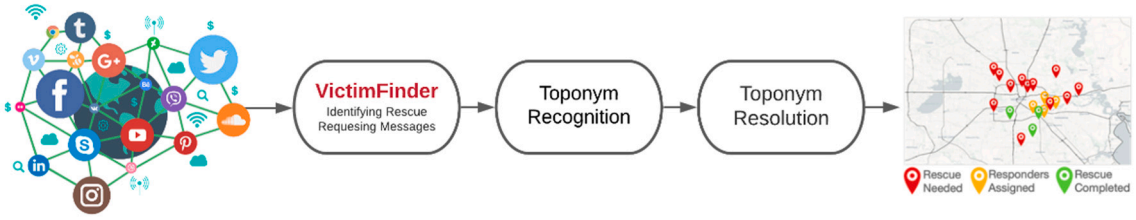


Fig. 1. The workflow of collecting rescue requests from social media.

cycle (Wang, Lam, Zou, and Mihunov, 2021; Wang, Singh, Tang, and Dai, 2017; Zou et al., 2019). The observed patterns offer a deeper understanding of social and geographical disparities in disaster-related social media use, which could help develop pathways to enhance resilience to hurricanes.

2.2. NLP for social media analysis

In recent years, NLP techniques have been used in multiple branches of social media analysis. For instance, NLP was applied in sentiment analysis to detect and extract reactions and opinions toward a given topic in social media, e.g., determining local reactions to disasters to improve emergency management (Beigi, Hu, Maciejewski, and Liu, 2016). A Neural Network structured model has been established to tackle multiclass classification among five different topics to gain cross-disaster situation awareness (Yu, Huang, Qin, Scheele, and Yang, 2019). A two-step approach that fuses a Neural Network binary classifier to detect disaster-related tweets and a Latent Dirichlet Allocation (LDA) method to extract fine-grained categories such as disrupted service and damaged facilities has been developed to analyze disaster impact (Sit, Koylu, and Demir, 2019). However, finding victim information from tweets is a topic that has rarely been visited by scholars. It could be a classification problem that requires natural language understanding. Leveraging NLP models for this task involves two steps: (***a***) text representation to map tweets into higher dimensions of matrices or vectors, and (***b***) training machine learning algorithms with the derived tweet matrices or vectors to perform classification (Xing, Pei, and Keogh, 2010). The first step comprises word tokenization, vectorization and word embedding, the history of which can stretch as far back as the application of the one-hot encoding approach in 2012 (Harris & Harris, 2012). Since 2013, numerous NLP text representation models relying on co-occurrence frequencies of words have been developed, such as the continuous bag-of-words model (Mikolov, Sutskever, Chen, Corrado, and Dean, 2013). These algorithms are confined because they lack high-level symbolic capabilities, namely, manipulating recursive and constituent structures, representing abstract concepts, lexical and semantic access, and episodic memories (Cambria and White, 2014).

Starting from the introduction and development of neural network-based NLP models (Bengio, Ducharme, Vincent, and Janvin, 2003), the boundaries of the above limitations have been pushed forward. The Recurrent Neural Network (RNN) and its modified versions consider the context of where the words appear and have achieved significant performance improvements in text representation and classification tasks compared with the one-hot encoding and frequency-based approaches (Peters et al., 2018).

However, the classification accuracy of RNN-based models ranges from 0.6 to 0.85 in classifying social media texts (Li, Li, and Zhu, 2016; Lee and Dernoncourt, 2016; Wang et al., 2017), which are insufficient to support their applications in life-concerning events such as victim finding. One bottleneck in RNN-based models is that only the output layer of the neural network in text representations is input to the downstream classification tasks (Vaswani et al., 2017). Consequently, the information contained in the initial part of the neural network sequence might vanish. Recently, this problem has been tackled by replacing RNN models with attention-based models (Vaswani et al., 2017), which consider the information contained in all hidden layers. One typical example of the attention-based NLP models is BERT.

The newest NLP technology brings about novel models that render better performance in many tasks but have rarely been applied in classifying rescue requesting tweets. Whether incorporating advanced NLP models could generate higher efficacy in identifying rescue request messages is unknown and solicits further investigations.

2.3. Hurricane Harvey

Hurricane Harvey was developed from a tropical wave on August 17th, 2017, and grew into a Hurricane when hitting the north of Colombia. On August 25th, 2017, Harvey made its first landfall in the United States as a category-4 hurricane on coastal Texas, followed by a few landfalls in Texas and Louisiana (Fig. 2). It caused at least 106 confirmed deaths in the United States and an estimated total of 125 billion dollars in damage (Feltgen, 2018). Harvey is the second-costliest natural disaster recorded in Texas, only after the 2021 Winter Storm Uri, which disabled the entire state's power grid (Ivanova, 2021). Most of the damages were caused by the catastrophic flooding in the Houston metropolitan area and Southeast Texas triggered by the unprecedented heavy rainfall with the peak accumulation reaching 153.87 cm. The flood inundated hundreds of thousands of homes, displacing over 30,000 people and prompting no less than 17,000 rescues (Blake and Zelinsky, 2018).

Several scholars have explored the emerging use of social media in disaster rescue during Hurricane Harvey. For example, user-friendly web applications have been developed and evaluated to detect and locate real time events based on geo-tagged streaming tweets by a novel cross-modal authority measure (Zhang et al., 2018). Victims and volunteers can be identified with a trained classifier whilst a hybrid scheduling logic is applied to ensure the most effective rescue work (Yang, Nguyen, Stuve, Cao, and Jin, 2017). Surveys have been conducted to investigate users who tweeted for help to reinforce future rescue operations and to understand how Twitter usage reshapes disaster rescue activities (Mihunov et al., 2020). An effective toponym extraction tool, NeuroTPR, has been trained to accurately extract locations from texts to determine victims' addresses in social media-based disaster rescue missions (Wang, Hu, and Joseph, 2020).

The aforementioned examples also reveal the challenges in using social media for disaster rescue (Mihunov et al., 2020). People typed their messages differently to call for help because no official standards were composed for requesting rescue on social media. Thus, volunteers manually detected, comprehended, and processed enormous social media data during hazards to pinpoint rescue-related messages and locate victims. This manual process requires intensive human labor and time to gather and organize information swiftly. More in-depth on developing automated tools for collecting and processing rescue request messages on social media is urgently needed.

3. Methodology

3.1. VictimFinder model architectures

The architecture of *VictimFinder* models consists of a pretrained model layer and a classifier layer, as delineated in Fig. 3. The pretrained model learns the general language representations from existing corpus while the classifier targeted at specific downstream tasks. In this way, we profit from the knowledge captured by large and complicated pretrained models and usher it to the goal of detecting rescue request tweets. Tweets were initially converted to vectors through a tokenization and vectorization process, which splits tweets into lists of tokens and represents each token by a pre-defined identification number. Then, through the language model pretrained on large amounts of texts, preprocessed tweets can be encoded to meaningful embeddings that capture the semantics of words or sentences. Finally, each tweet embedding is fed into classifiers to detect if the user is requesting rescue or not.

There are two common methods to train *VictimFinder* models based on the proposed architecture: fine-tuning and feature-based approaches (Devlin et al., 2019). The fine-tuning method calculates and modifies the parameter in both pretrained models and classifiers while training on a specific downstream task. In feature-based approaches, the parameters within the pretrained language model remain static, and only the classifier parameters are modified during training. Thus, compared to fine-tuning strategies, fewer parameters need to be trained in feature-based approaches. This architecture is highly expandable and allows quick alternation between numerous model types and training routines.

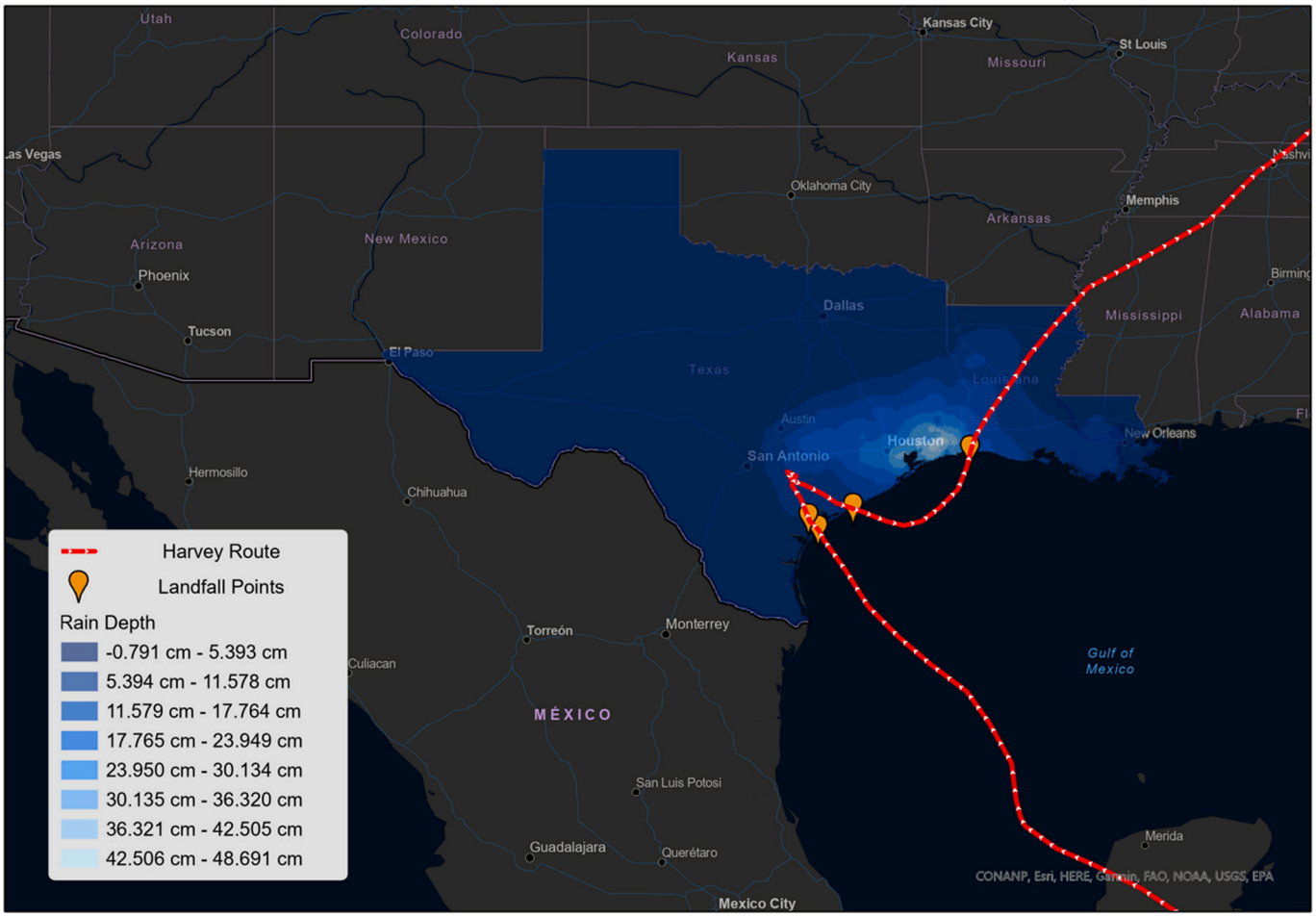


Fig. 2. The route and rain depth of 2017 Hurricane Harvey.

This study built ten *VictimFinder* models using fine-tuning or feature-based approaches based on seven pretrained language models, including Global Vectors for Word Representation (GloVe), Embeddings from Language Models (ELMo), BERT, RoBERTa, DistilBERT, ALBERT, and XLNet (Table 1). GloVe and ELMo are milestone NLP models and rendered state-of-the-art performance in several NLP tasks when they were released. They served as baseline models in this research. The original BERT and its modified versions (RoBERTa, DistilBERT, and ALBERT) were tested to compare the performance of identifying rescue requests based on different existing BERT models. XLNet, which was proposed to overcome the limitations of BERT-based language models, was also included to investigate if the more complicated language model could outperform BERT and BERT-based models in building *VictimFinder*.

GloVe and ELMo were coupled with a Transformers classifier, which is adopted from the Attention mechanism to capture more information from natural language (Vaswani et al., 2017). They are trained through the feature-based approach (models 1&2 in Table 1). The default BERT, modified BERT, and XLNet models were trained through a fine-tuning approach (models 3–7).

In addition, we proposed three novel designed customized BERT-based models by placing a nonlinear multi-layer neural network, a Convolution Neural Network (CNN; O'Shea and Nash, 2015), and a Long-Short Term Memory (LSTM; Greff, Srivastava, Koutník, Steunebrink, and Schmidhuber, 2017) above the BERT model to process classification tasks (models 8–10). Divergent from default BERT with linear classifier, the BERT-Nonlinear model attempts to boost the performance with a more complicated classifier. The BERT-CNN and BERT-LSTM models try to harness the information captured by BERT exhaustively by incorporating all hidden states parameters to tackle information vanishing problems. The three customized models were trained using the feature-based approach, meaning tweets in the training dataset were first encoded to embeddings using the default BERT model and then input to neural network-based classifiers to optimize their parameters. Compared to fine-tuned BERT-based models, customized BERT-based models take advantage of the pretrained embeddings and are expected to detect higher dimensional features hidden in the embedded texts with a limited training dataset.

3.2. Selected pretrained models

3.2.1. GloVe

Representing words by real-valued vectors is commonly used in the domain of NLP that the vectors can be input as features in various downstream applications and text categorization (Tellex, Katz, Lin, Fernandes, and Marton, 2003; Sebastiani, 2002; Turian, Ratinov, and Bengio, 2010; Socher, Bauer, Manning, and Ng, 2013). There are two major models of learning word representation, global matrix factorization and local context window methods, but either of the two models suffers from several drawbacks, e.g., performs poorly on word analogy or missing statistical information.

GloVe is a global log-bilinear regression model that combines the merits of these two types of word representation models. GloVe trains a specific weighted least squares model on the non-zero entries of a global word-word co-occurrence matrix, which tabulates the frequencies of words with another given word in a corpus. Five corpora of different sizes are utilized to train the model. It generates word vectors with meaningful substructure that leads to better performance on several

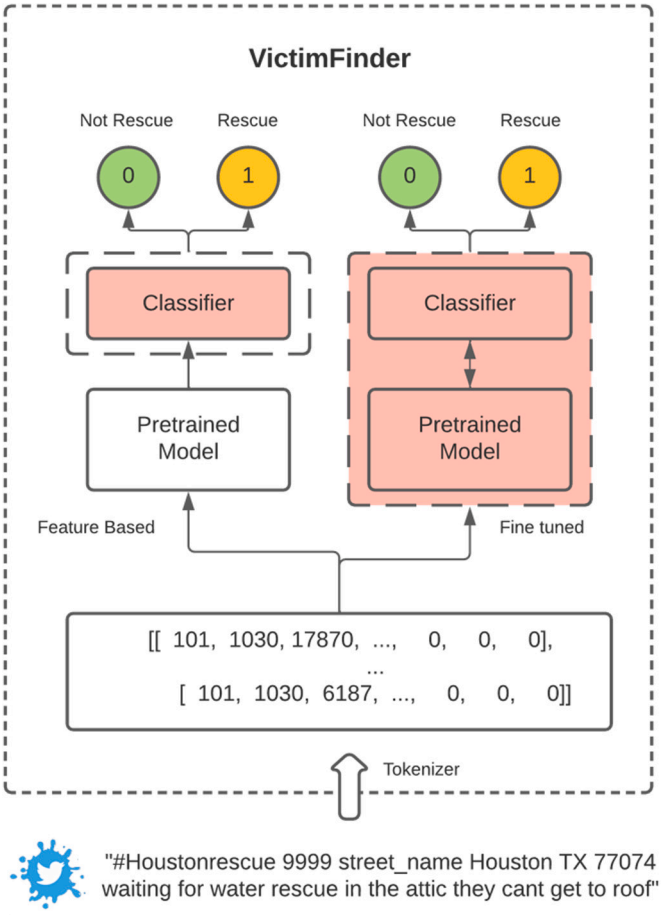


Fig. 3. The architecture of NLP-based *VictimFinder* models.

Table 1
The developed *VictimFinder* models and training methods.

No.	VictimFinder models	Pretrained models	Classifiers	Training methods
1	Glove-Transformers	GloVe	Transformers	Feature-based
2	ELMo-Transformers	ELMo	Transformers	Feature-based
3	BERT-Linear	BERT	Linear	Fine-tuning
4	RoBERTa-Linear	RoBERTa	Linear	Fine-tuning
5	DistilBERT-Linear	DistilBERT	Linear	Fine-tuning
6	ALBERT-Linear	ALBERT	Linear	Fine-tuning
7	XLNet-Linear	XLNet	Linear	Fine-tuning
8	BERT-Nonlinear	BERT	Nonlinear	Feature-based
9	BERT-LSTM	BERT	LSTM	Feature-based
10	BERT-CNN	BERT	CNN	Feature-based

tasks than continuous bag-of-words and skip gram models, such as word analogy, named entity recognition, and word similarity tasks (Pennington et al., 2014).

3.2.2. ELMo

Instead of relying on word co-occurrence counts, ELMo leverages vectors derived from a bidirectional LSTM which is trained on a huge text corpus with a coupled language model objective (Peters et al., 2018). This is a novel deep contextualized word representation that takes consideration of both the word uses and how these uses change across linguistic contexts. The difference between ELMo and traditional word type embeddings, e.g., word2vec and GloVe, is that every token is derived from a function of the entire sentence input. Therefore, the same word can have distinguished representation vectors under different contexts. Moreover, most LSTM-based language models at that phase in time run only in the forward direction, while ELMo takes advantage of a bidirectional approach in which the model also runs over the sequence in the reverse direction and both the forward and backward language models are combined.

The word vectors are calculated on top of a two-layer bidirectional language model (biLM). These two layers are stacked together and each one has 2 passes – a forward pass and a backward pass. The words are tokenized using a convolutional neural network and are input into the biLM. The output of the forward and backward passes is concatenated to form the intermediate word vectors and are input into the next hidden layer. The final output of ELMo can easily be used upon other existing models for specific downstream tasks. Its application has brought noticeable improvement to 6 challenging NLP tasks (question answering, textual entailment, semantic role labeling, coreference resolution, named entity extraction, and sentiment analysis) compared to preceding models, including GloVe.

3.2.3. BERT

Feature-based approaches are applied to the former language models that achieve state-of-the-art performance with a General Language Understanding Evaluation (GLUE) benchmark score of 71.0. However, the techniques leveraged by them have limitations. For example, ELMo is based on a bidirectional LSTM architecture rather than a Transformer architecture, in which valuable information in the inchoate hidden layers may vanish when the recurrent network goes deeper. BERT which is designed to pretrain deep bidirectional representations using unlabeled text has been introduced (Devlin et al., 2019) settling such limitations by applying bidirectional Transformer, which is an attention mechanism that learns contextual relations between words during fine-tuning. The model architecture is shown in Fig. 4.

Transformer contains an encoder mechanism and decoder mechanism. Since the aim of BERT in this study is to generate a language model, only the encoder part of Transformer is required. Unlike many bidirectional language models, in which the contextual representation of every token is the concatenation of the forward and backward representations, the Transformer encoder reads the entire sequence of words at once. In this sense, it is considered bidirectional or non-directional, which enables the model to learn the context of a word based on all of its surroundings. To implement such bidirectional approaches, BERT leverages two training strategies. First, BERT proposes a Masked Language Model (MLM) inspired by the Cloze task (Taylor, 1953) in which 15% of the input tokens are masked by a special label [mask] at random and then predict those masked tokens. Second, BERT introduces Next Sentence Prediction (NSP) to the training process. The model takes in pairs of sentences as input and attempts to identify if the second sentence within the input pair is the subsequent one in the

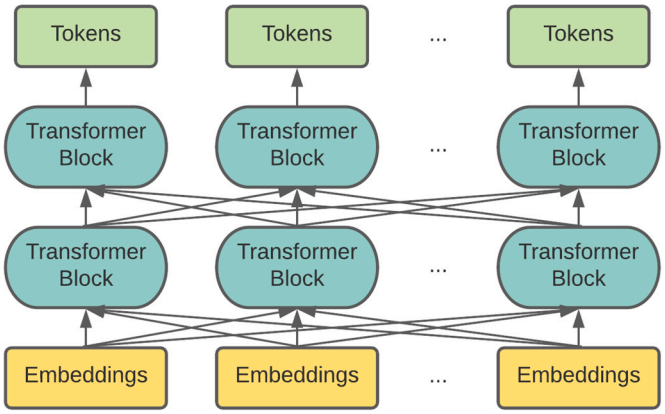


Fig. 4. Architecture of BERT.

original document.

BERT can be fine-tuned with a simple output layer to generate state-of-the-art performance throughout a number of NLP tasks. Under the GLUE benchmark, BERT outperforms ELMo by approximately 7% (Devlin et al., 2019).

3.2.4. RoBERTa

Since BERT started to prevail in the NLP world, research on developing BERT-based models has emerged. RoBERTa, for example, is a novel and improved recipe for training BERT models, which can match or surpass many post-BERT methods (Liu et al., 2019). It was proposed after meticulously measuring the impact of key hyperparameters and finding that BERT was undertrained. RoBERTa optimizes the training process of BERT in the following ways. First, RoBERTa offers the training process more time, with bigger batch sizes and more data. Second, the next sentence prediction objective in BERT is removed in RoBERTa. Third, RoBERTa trains the model with a longer sequence. Last, the patterns of the masked language model are altered dynamically. Compared to BERT, RoBERTa achieves a leading score of 88.5 on the public GLUE benchmark result board over numerous tasks.

3.2.5. DistilBERT

DistilBERT is a smaller pretrained model on the basis of BERT using knowledge distillation. This process is a compression strategy to train a compact model, called the student, capable of duplicating the performance of a larger model or an ensemble of models. DistilBERT is built on the same general structure of BERT with token-type embeddings while removing the poolers and reducing the number of layers. It also takes advantage of the training objectives recommended by RoBERTa. Experiment results show that DistilBERT significantly reduces the model size and accelerates the training and predicting speed while maintaining most of the performance (Sanh, Debut, Chaumond and Wolf, 2020).

3.2.6. ALBERT

ALBERT is another simplified version of BERT that is also constructed on the transformer encoder architecture. It utilizes two parameter reduction techniques to achieve faster training speed at lower memory consumption. The first method is factorized embedding parameterization, which projects the one-hot vectors into a lower dimensional space before projecting them into the hidden space during tokenization. The second one is cross-layer parameter sharing, a common technique aiming to improve parameter efficiency. There are several strategies of parameter sharing, among which ALBERT chooses to share parameters across layers. A self-supervised loss focusing on modeling inter-sentence coherence has been applied which contributes consistently to the performance of downstream tasks with multi-sentence inputs. Empirical evidence shows that ALBERT establishes new state-of-the-art results on GLUE benchmark with an average score of 88.7 (Lan et al., 2020).

3.2.7. XLNet

BERT implements its bidirectional architecture by corrupting the input text with a special label [mask] which neglects the dependencies between the masked locations. Moreover, the data used for fine-tuning BERT do not contain this [mask] label, leading to a pretrain-finetune discrepancy. In light of this, XLNet, a generalized autoregressive pretraining method, has been proposed. XLNet enables training bidirectional contexts by maximizing the expected probability over all permutations of the factorization order, and its autoregressive formulation helps overcome the limitations of BERT (Yang et al., 2020). In addition, XLNet integrates ideas from an advanced autoregressive model, Transformer-XL, and designed a two-stream attention mechanism. Under similar experiment settings, XLNet renders better performances compared to BERT across 20 common NLP tasks (Yang et al., 2020).

4. Experiments

4.1. Training dataset

The training dataset was extracted from Twitter data purchased from the Twitter company. Harvey-related Twitter data during August 25th–31st, 2017, from Harvey's first landfall to when it weakened to a storm, were collected through Twitter's enterprise Application Programming Interface (API), which provides the full historical Twitter data through keyword search or criteria-based search, such as location-based and user-based search (https://developer.twitter.com/en/docs/twitter-api/enterprise). We used a list of case-insensitive keywords to identify an initial collection of Harvey or rescue-related tweets: [*harvey, hurricane, storm, flood, houston, txtf* (Texas Task Force)*, coast guard, uscg* (U.S. Coast Guard)*, houstonpolice* (Houston Police Department)*, cajun navy, fema* (Federal Emergency Management Agency)*, rescue*]. Every tweet containing at least one of the keywords was retrieved, resulting in 25 million tweets in the initial collection. Then we chose original English tweets containing the 5-digit zip code of coastal Texas as potential rescue request tweets, which returned 3191 unduplicated tweets.

We manually labeled each of the 3191 tweets by four questions to prepare the training database. The questions were designed based on a suggested rescue request method posted by a volunteer organization's Twitter account (@HarveyRescue) during Hurricane Harvey. The method recommends Twitter users asking for rescue to post tweets with complete address, number of people who need help, phone number, and other special needs while hashtagging #HarveySOS and mentioning the account. Each tweet was binarily categorized based on four questions: is the tweet asking for help (label: help)? Does the tweet provide a full address (fullAddress)? Does the tweet mention the demographic or health-related information of victims, e.g., gender, age, physical conditions, and special needs (victimInfo)? Does the tweet describe the hazard situations, e.g., flooded water levels (hazardSituation)? We labeled the tweet as positive for each question if the answer is yes and negative for no. Two student workers were hired to annotate the tweets with timely communication to address the disagreements on label types. The tagged data were also scrutinized by the author to ensure consistency in tagging. Five examples of the labeled data are shown in Table 2. Place names and street names appearing in the sample data shown in the table were replaced by 'place_name' and 'street_name'. Street and place numbers were replaced by random numbers.

The distributions of the four types of labels are shown in Fig. 5(a). Among the 3191 manually labeled tweets, 1935 (60.64%) were seeking help, 2003 (62.77%) provided full addresses, 1278 (40.05%) mentioned victims' conditions and special needs, and 57 (1.79%) described their hazard situations. Addresses in the rescue request tweets were further extracted using a customized regular expression and geocoded through the Google Geocoding API. Fig. 5(b) reveals the spatial hot spots of the rescue requests during Harvey extracted from the training dataset. The temporal patterns of the number of rescue request tweets within the dataset is shown in Fig. 5(c).

In this study, only the first three labels (help, fullAddress, and victimInfo in Fig. 5(a)) were used because their positive and negative labels' distributions are ideal for model training. All labeled data were first input into a text cleaning process where all texts were lowercased, and non-ASCII letters were removed. Punctuation mark removal and lemmatization were performed before the experiment.

4.2. Implementation

Table 3 lists the initial pre-trained parameters and concise descriptions for each *VictimFinder* model. For example, the pretrained model for BERT is designed with 12 layers. The size of the hidden layer is 768 and the number of self-attention head is 12. This model has 110 million parameters and is pre-trained using unlabeled text from

Table 2
Examples of manually labeled tweets.

Text	help	fullAddress	victimInfo	hazardSituation
#Houstonrescue 9999 street_name Houston TX 77074 waiting for water rescue in the attic they cant get to roof	1	1	0	0
Rescue Needed ForemanFamily of house flooding in last hours 11,870 place_name 77,044	1	1	1	0
@user_name55555 praying you & your family will be out of harms way Hurricane Harvey is something serious	0	0	0	0
Harvey is expected to make landfall as Category hurricane Text HARVEY to 77,453 to receive alerts on the massive Texas storm	0	0	0	0
@user_name My friend her roommate and their dogs in attic house flooding street_name TX 77539 #dickinson #galveston #houstonflood	1	1	1	0

BooksCorpus (800 million words) and English Wikipedia (2500 million words) with masked language model and next sentence prediction method as mentioned in Section 3.2.

A dummy classifier, which aims to authenticate the performance of the models tested and detects the possible bias exist in the testing datasets, is included in this experiment. The stratified strategy is applied to the dummy classifier, which generates predictions by respecting the training set's class distribution.

Before inputting to *VictimFinder* models, tweets were initially tokenized and vectorized through the BERT tokenizer (Devlin et al., 2019). The output embeddings of GloVe and ELMo were input into a two-head, one-layer Transformer with average pooling for classification. The input layer size was 512, and the dropout rate was set to 0.3. Models 3–7 were linked to a simple linear classifier to perform softmax, which normalizes the output to a probability distribution over predicted output classes.

The three customized feature-based models were implemented by running the default BERT model first, as shown in Table 3 and Fig. 6. For the BERT-Nonlinear model, the final hidden state of the default BERT model was transferred to a neural network with three fully connected layers of 256 nodes. Leaky Rectified Linear Unit (LeakyReLU) was leveraged as the activation function (negative slope = 0.01) of the first two layers and the final layer output the classification results by performing softmax function. For the BERT-LSTM model, a bidirectional LSTM was constructed and placed on top of BERT, and the last hidden state of LSTM was linked to a fully connected layer to perform softmax. The BERT-CNN model used hidden states from all layers of BERT. A convolution kernel was performed with 16 filters on the hidden states. The output data were concatenated to form a channel with 192 nodes. A one-dimensional vector was formed through max pooling and linked to a fully connected layer to perform softmax.

The labeled dataset was randomly separated as training data, cross-validation data, and testing data with a ratio of 8:1:1. Random state was stored to ensure the data used for each model evaluation process were the same. The number of maximum training epochs was set to 40 with early stopping. As suggested in previous literature, the learning rate for non-BERT models was set to 0.001, and 2e-5 for BERT-based models (Devlin et al., 2019). We chose the AdamW algorithm as the optimizer and cross entropy as the loss function.

All models were established using Python and relevant packages. Dummy classifiers were implemented using the ScikitLearn library with the stratified strategy. Models based on GloVe and ELMo were implemented using PyTorch. The BERT-based models were implemented based on the huggingface Transformer library (https://huggingface.co/transformers/). Fine-tuning models were implemented by using huggingface sequence classification functions. Customized BERT-based models were built using PyTorch functions.

4.3. Evaluation criteria

The three most used metrics, precision, recall, and F1-score (Eqs. 3–5), were utilized to evaluate the performance and bias of *VictimFinder* models. Precision measures the percentage of correctly identified tweets (noted as True Positives, TP) among all the positive tweets detected by the model, which combines both TP and False Positives (FP). Recall measures the percentage of correctly identified tweets among all ground truth, which is the combination of TP and False Negatives (FN). F1-score is the harmonic mean of precision and recall, providing a comprehensive metric to evaluate model performance. GloVe, which has the longest history, was selected as the baseline model for accuracy evaluation.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{–}score = 2*\frac{Precision*Recall}{Precision + Recall} \tag{5}$$

Considering that rescue tweets detection should be timely so that related agencies could provide immediate assistance during disasters, we also measured the computational costs by comparing the training time and predicting time of each selected model. Fine-tuning BERT was selected as the baseline model to present a clearer view of the time cost for training and predicting.

Each model was applied in three independent classification tasks. The standard deviation is calculated to find the most versatile model. To normalize the discrepancy between different models, we leveraged a normalization mapping considering the conception of the softmax function. A novel stability metric, the Normalized Model Stability Index (NMSI; Eq. 6), where all the output number of the function sums to 1 and the order is preserved, was introduced.

$$NMSI = 1 - \frac{1}{\sum_{k=1}^{k} e^{\sigma_k}} \bullet e^{n\sigma_k} \tag{6}$$

In Eq. 6, the number of selected models is noted as k ($k = 3$ in this study), σ stands for the standard deviation of F1-score, and n is the augment factor and was set to 100 in this case.

5. Results

Evaluation metrics of model performances are summarized in Tables 4 and 5. According to Table 4, the F1-score of GloVe and ELMo ranged from 0.747 to 0.858, while the BERT-based models ranged from 0.834 to 0.919. XLNet, which takes the asset from the Transformer structure and aims to optimize BERT, ranged from 0.824 to 0.905. All BERT-based models generally outperformed the baseline model by approximately 10%. The result verifies the hypothesis that the state-of-the-art model performs better on rescue tweet classification tasks. This is largely due to BERT's word piece embedding, masked language modeling, and the application of a bidirectional Transformer (Devlin et al., 2019), leading to a better understanding of natural language by computer programs.

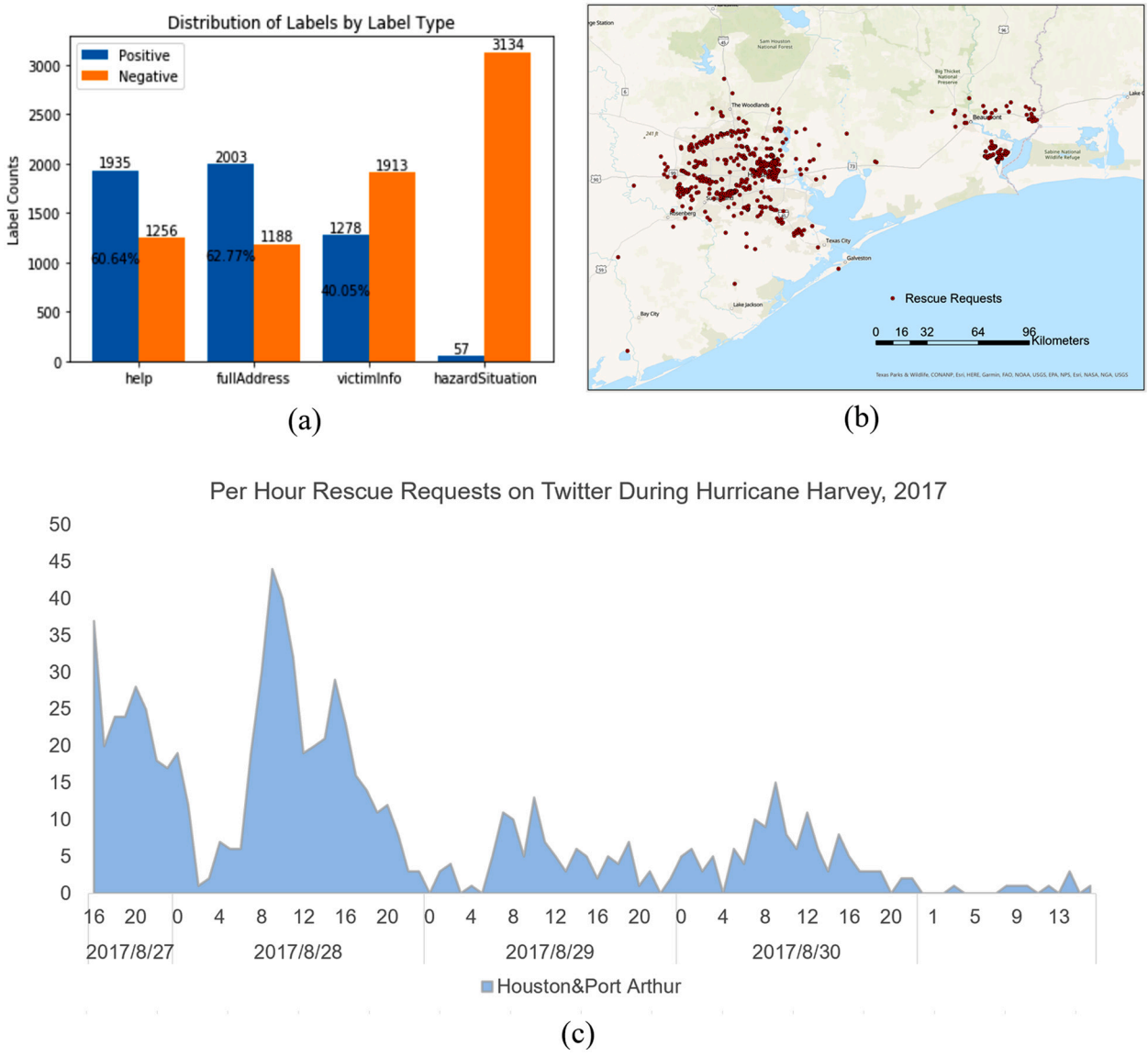


Fig. 5. The distribution of labels in the training dataset and their spatial-temporal patterns.

Table 3
Pre-trained parameters for developed tweet classification models.

No.	VictimFinder models	Pre-trained parameters	Description
–	Dummy	–	–
1	Glove-Transformers	glove.twitter.27B-100d	Train with Twitter data, 2B tweets, 27B tokens, 1.2M vocab, uncased, 100d vectors
2	ELMo-Transformers	elmo_2x2048_256_2048cnn_1xhighway_weights. hdf5	Parameters 28.0 (Millions) LSTM Hidden Size/Output size 2048/256, Highway Layers>1
3	BERT-Linear	bert-base-uncased	12-layer, 768-hidden, 12-heads, 110M parameters, trained on lower-cased English text
4	RoBERTa-Linear	roberta-base	12-layer, 768-hidden, 12-heads, 125M parameters; using the BERT-base architecture
5	DistilBERT-Linear	distilbert-base-uncased	6-layer, 768-hidden, 12-heads, 66M parameters; distilled from BERT model bert-base-uncased
6	ALBERT-Linear	albert-base-v1	12 repeating layers, 128 embedding, 768-hidden, 12-heads, 11M parameters
7	XLNet-Linear	xlnet-base-cased	12-layer, 768-hidden, 12-heads, 110M parameters; XLNet English model
8	BERT-Nonlinear	Same as Model 3	Same as Model 3
9	BERT-LSTM	Same as Model 3	Same as Model 3
10	BERT-CNN	Same as Model 3	Same as Model 3

Certain models perform better than the rest for specific tasks. In identifying rescue requesting tweets, feature-based BERT with CNN classification head performed the best with a 0.919 F1-score. Fine-tuning BERT with linear classification head had the best performance for detecting full address information within the text with an F1-score of 0.913, while feature-based BERT with LSTM classifier outperformed the rest in distinguishing tweet with or without victim information with an F1-score of 0.856. However, no gigantic gap across the performance of

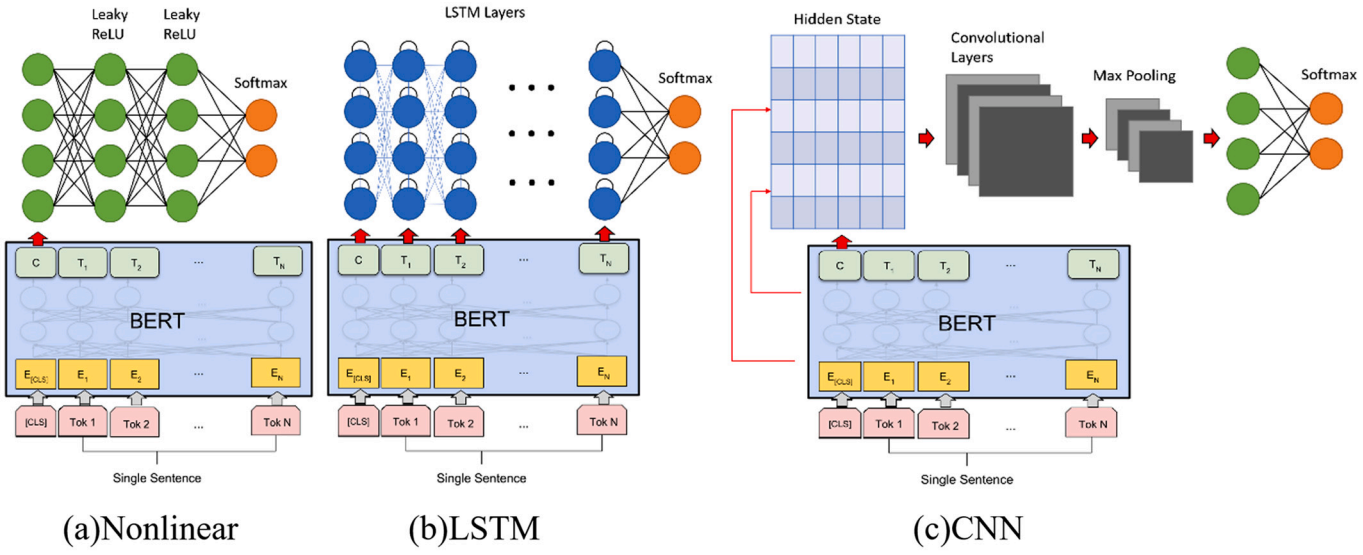


Fig. 6. Structures of the three customized BERT models.

Table 4
Precision (P), Recall (R), and F1-score of each model.

VictimFinder models	Label-"help"			Label-"fullAddress"			Label-"victimInfo"		
	P	R	F1	P	R	F1	P	R	F1
Dummy	0.642	0.618	0.630	0.628	0.631	0.629	0.666	0.641	0.653
GloVe-Transformers	0.782	0.850	0.831	0.756	0.845	0.813	0.806	0.696	0.747
ELMo-Transformers	0.779	**0.963**	0.846	0.800	**0.963**	0.858	0.812	0.859	0.790
BERT-Linear	0.884	0.918	0.909	**0.889**	0.933	**0.913**	0.869	0.825	0.838
Roberta-Linear	0.886	0.918	0.910	0.878	0.918	0.904	0.873	0.871	0.850
DistilBERT-Linear	0.881	0.903	0.905	0.870	0.940	0.901	0.862	0.840	0.834
XLNet-Linear	0.862	0.915	0.893	0.873	0.963	0.905	0.840	**0.905**	0.824
ALBERT-Linear	0.854	0.875	0.883	0.886	0.938	0.912	0.870	0.840	0.842
BERT-LSTM	0.878	0.913	0.904	0.873	0.940	0.903	**0.881**	0.856	**0.856**
BERT-CNN	**0.897**	0.933	**0.919**	0.873	0.940	0.903	0.867	0.817	0.835
BERT-Nonlinear	0.881	0.900	0.905	0.883	0.920	0.908	0.856	0.882	0.835

Table 5
Computation cost and model stability.

VictimFinder models	Training time ratio	Predicting time ratio	NMSI
GloVe-Transformer	**0.257**×	**0.006**×	0.858
ELMo-Transformer	0.278×	0.189×	0.927
BERT-Linear	1.000×	1.000×	0.878
Roberta-Linear	1.019×	0.940×	0.942
DistilBERT-Linear	0.521×	0.494×	0.900
XLNet-Linear	**1.453**×	**2.290**×	0.860
Albert-Linear	0.834×	0.997×	0.933
BERT-LSTM	1.021×	1.004×	**0.963**
BERT-CNN	1.022×	1.006×	**0.853**
BERT-Nonlinear	1.023×	1.005×	0.886

all selected Transformer-based models was observed. BERT models customized with more complicated encoders may have better results but only marginally, possibly due to training data deficiency. This can be further proven by the fact that models with larger numbers of parameters, such as XLNet and RoBERTa, performed worse than BERT-based models. ELMo had a noticeably higher recall for the first two tasks which indicates that it has a tendency of labeling tweets as positive samples. This is helpful for situations when fewer positive samples are allowed to be missed out.

According to Table 5, the training time of models using GloVe and ELMo were 0.257 and 0.278 times of BERT-based models (1.000×), and the predicting time was 0.006 times and 0.189 times of BERT-based models (1.000×), respectively. The XLNet-linear model, the most complicated model in this study, cost nearly 50% extra training time than the BERT-linear model, and the predicting time was doubled. This corresponds with the fact that Transformer models are massive and contain millions of parameters, making them computationally more expensive. The NMSI ranged from 0.853–0.963. The BERT with LSTM classifier performed the most stably across three classification tasks with an NMSI value of 0.963, which is 12.2% better than the baseline model (BERT-Linear). However, BERT with CNN classifier performed inferiorly regarding model stability with the lowest NMSI of 0.853. This might be caused by the hidden feature disparities between the embedded texts of help request, address, and victim information.

Considering both Table 4 and Table 5, the customized BERT-LSTM model has performed most stably upon three classification tasks. This means the BERT-LSTM structure could be more trustworthy when applied to other tasks such as cross-event prediction. Regarding time efficiency, the actual difference may not be significant between the maximum number 2.290× and the minimum number 0.006× when running the model on tiny testing datasets. However, in disaster events, the model will be utilized to predict real-time Twitter streaming data, which are both high in flow and volume. The deviations in computational overheads mean a gigantic gap in the time consumed in this scenario. Consequently, it is reasonable to consider compensating efficiency with accuracy. The DistilBERT-Linear model renders similar performance than other models but is much more efficient, making it the optimal choice for rescue tweet detection tasks accounting for both accuracy and efficiency.

6. Conclusion

This paper utilized different NLP models for sequence classification on harvesting rescue requests from Twitter data. The objective was to examine if BERT-based models can achieve better performance than models based on milestone NLP algorithms in rescue request tweet classifications. Experiment results show that all BERT-based models outperformed the baseline model with a limited amount of training data. BERT models with customized classification heads led to significant improvement in performance compared to the baseline model. In terms of identifying help requesting tweets, the best performer, BERT with CNN classifier, obtained a 0.919 F1-socre, which is 10.6% better than the baseline model. BERT with LSTM classifier performed most stably across all three classification tasks. DistilBERT, considering both model performance and efficiency, could be the most appropriate model for rescue request tweets detection in disaster response. ELMo, which processes data extremely quickly with very high recall, can be applied to situations where fewer rescue request tweets are expected to be missed out.

This study contributes both scientifically and practically in several aspects. Initially, by constructing a labeled dataset, we make training victim-finding models possible and bring state-of-the-art NLP technology to this domain of study. Furthermore, this research provides valuable insight into which NLP model should be selected to categorize rescue request tweets based on experiments with real rescue request Twitter data during Hurricane Harvey. Finally, the result of the experiments acts as the cornerstone of future victim finding applications. Web applications can be developed, and the optimal model can also be incorporated into GIS tools for displaying near real-time rescue request locations to which emergency responders and volunteers can refer to send help.

However, several issues may arise due to the nature of big social media data analysis and some limitations of this research. Such issues should not be ignored while translating research results into practice. First, people using languages other than English on social media cannot leverage the benefit of these research results. One way to solve the problem is to train a corresponding model for the target language or scrutinize whether a unified model can render reliable performance across numerous languages. Second, bias in Twitter datasets is a common non-negligible issue. This approach tends to help those who use social media more often, and those groups of people may not be the ones who demand help the most during disasters. Uneven usage of social media may lead to biased consequences. Moreover, social media posts suffer from locational bias, temporal bias, and reliability issues. Such issues should be considered while further analyzing the spatiotemporal patterns of the identified tweets for detecting vulnerable communities or assessing disaster damages. Third, the models are trained and tested with tweets from a single event, Hurricane Harvey, that may affect the generalizability of the model. Incorporating Twitter data from other events of the same disaster type will reinforce the robustness of the model developed. Fourth, a large amount of sensitive data containing users' privacy can be extracted from social media by applications built upon such classifiers. Meticulous consideration should take place to decide who should have access to these applications and databases, or it might cause privacy issues. Fifth, the best model mentioned in this research offers an F1-score of 0.919, which is a promising result. However, if more weight is put on social media rescue requests, the misclassified requests may mean lives lost without scrutiny. Continuing to assemble disaster rescue information from multiple spectrums and practice social media data as a supplementary source for finding and rescuing disaster victims will solve this limitation.

Future studies can be conducted in the following directions to address the limitations of the study and advance social media use in disaster rescue. The model performance can be further optimized by fine-tuning the hyperparameters and utilizing larger training datasets. Data from other events of the same disaster type can be collected and utilized to test and foster the generalizability of the model. Second, a rescue tweets detection and analysis pipeline can be constructed with the optimized model by adding advanced geoparsers and geocoding techniques. Through applying the pipeline and spatial analysis, valuable information and discovery regarding overlooked communities and limitations of traditional disaster management can be extracted from massive tweets to enhance disaster mitigation and preparedness. Third, the scope of rescue request detection should not be limited to NLP text classification only. The accuracy of detecting rescue request tweets can be improved by processing the images posted along with the text data. In addition, being inspired by the concept raised by SocioDim, which is a classification framework based on network structure to capture interaction patterns (Tang and Liu, 2011), and Global Consistency Maximization, which is a link-based classification model to identify opinions (Li et al., 2016), network-based approach might be a feasible shortcut toward enhanced rescue request discovery. Fourth, interesting patterns can be recognized through investigating the spatial, temporal, textual, and diffusion characteristics of rescue request tweets. The results can inform disaster response, rescue operation, and damage estimation in future events. Fifth, Hurricane Harvey affected different social groups from various communities. The geographical features of rescue request tweets and the underlying socioeconomic characteristics of individuals and communities requiring additional assistance can be further evaluated.

Funding

This article is based on work supported by two research grants. One is from the U.S. National Science Foundation: Reducing the Human Impacts of Flash Floods - Development of Microdata and Causal Model to Inform Mitigation and Preparedness (Award No. 1931301). The other one is from the X-Grant program under the Texas A&M University President's Excellence Fund. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

CRediT authorship contribution statement

Bing Zhou: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft. **Lei Zou:** Conceptualization, Validation, Data curation, Resources, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing. **Ali Mostafavi:** Validation, Writing – review & editing. **Binbin Lin:** Software, Writing – review & editing. **Mingzheng Yang:** Writing – review & editing. **Nasir Gharaibeh:** Writing – review & editing. **Heng Cai:** Writing – review & editing. **Joynal Abedin:** Writing – review & editing. **Debayan Mandal:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

This article is based on work supported by two research grants. One is from the U.S. National Science Foundation: Reducing the Human Impacts of Flash Floods - Development of Microdata and Causal Model to Inform Mitigation and Preparedness (Award No. 1931301). The other one is from the X-Grant program under the Texas A&M University President's Excellence Fund. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

Data availability

Data will be made available on request.

References

Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. P. (2018). Social sensing of floods in the UK. *PLoS One, 13*(1), Article e0189327. https://doi.org/10.1371/journal.pone.0189327

Avvenuti, M., Cresci, S., La Polla, M. N., Marchetti, A., & Tesconi, M. (2014). Earthquake emergency management by social sensing. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)* (pp. 587–592). https://doi.org/10.1109/PerComW.2014.6815272

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. In W. Pedrycz, & S.-M. Chen (Eds.), *Studies in computational intelligence, 1860-949X: Volume 639. Sentiment analysis and ontology engineering: An environment of computational intelligence* (pp. 313–340). Springer. https://doi.org/10.1007/978-3-319-30319-2_13.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research, 3*, 1137–1155.

Blake, E. S., & Zelinsky, D. A. (2018). *Hurricane Harvey (no. AL092017; National Hurricane Center Tropical Cyclone Report)* (p. 77). National Hurricane Center. https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf.

Cambria, E., & White, B. (2014). Jumping NLP curves: a review of natural language processing research [review article]. *IEEE Computational Intelligence Magazine, 9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227

Cervone, G., Sava, E., Huang, Q., Schnebele, E., Harrison, J., & Waters, N. (2016). Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. *International Journal of Remote Sensing, 37*(1), 100–124. https://doi.org/10.1080/01431161.2015.1117684

Correll, R. M., Lam, N. S., Mihunov, V. V., Zou, L., & Cai, H. (2021). Economics over risk: flooding is not the only driving factor of migration considerations on a vulnerable coast. *Annals of the American Association of Geographers, 111*(1), 300–315.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint.* , Article arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Feltgen, D. (2018). *Harvey, Irma, Maria and Nate retired by the World Meteorological Organization*. National Oceanic and Atmospheric Administration. https://www.noaa.gov/media-release/harvey-irma-maria-and-nate-retired-by-world-meteorological-organization.

Gallagher, J. J. (2017). Hurricane Harvey wreaks historic devastation: by the numbers. In *ABC News*. Retrieved November 11, 2021, from https://abcnews.go.com/US/hurricane-harvey-wreaks-historic-devastation-numbers/story?id=49529063.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems, 28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Harris, D. M., & Harris, S. L. (2012). Chapter 2: Combinational logic design. *Digital Design and Computer Architecture,* (Second). Elsevier.

Ivanova, I. (2021). Texas winter storm costs could top $200 billion—More than hurricanes Harvey and Ike. Retrieved November 11, 2021, from https://www.cbsnews.com/news/texas-winter-storm-uri-costs/.

Jongman, B., Wagemaker, J., Romero, B. R., & De Perez, E. C. (2015). Early flood detection for rapid humanitarian response: harnessing near real-time satellite and twitter signals. *ISPRS International Journal of Geo-Information, 4*(4), 2246–2266. https://doi.org/10.3390/ijgi4042246

Kirilenko, A. P., & Stepchenkova, S. O. (2014). Public microblogging on climate change: one year of twitter worldwide. *Global Environmental Change, 26*, 171–182. https://doi.org/10.1016/j.gloenvcha.2014.02.008

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances, 2*(3), Article e1500779. https://doi.org/10.1126/sciadv.1500779

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint.* , Article arXiv:1909.11942. https://doi.org/10.48550/arXiv.1909.11942

Lee, J. Y., & Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint.* , Article arXiv:1603.03827. https://doi.org/10.48550/arXiv.1603.03827

Li, J., Li, X., & Zhu, B. (2016). User opinion classification in social media: a global consistency maximization approach. *Information & Management, 53*(8), 987–996. https://doi.org/10.1016/j.im.2016.06.004

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social sensing: a new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers, 105*(3), 512–530. https://doi.org/10.1080/00045608.2015.1018773

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., , … Stoyanov, V., et al. (2019). RoBERTa: a robustly pptimized BERT pretraining approach. *arXiv preprint.* , Article arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Mihunov, V. V., Lam, N. S. N., Zou, L., Rohli, R. V., Bushra, N., Reams, M. A., & Argote, J. E. (2018). Community resilience to drought hazard in the south-central United States. *Annals of the American Association of Geographers, 108*(3), 739–755. https://doi.org/10.1080/24694452.2017.1372177

Mihunov, V. V., Lam, N. S. N., Zou, L., Wang, Z., & Wang, K. (2020). Use of Twitter in disaster rescue: lessons learned from Hurricane Harvey. *International Journal of Digital Earth, 13*(12), 1454–1466. https://doi.org/10.1080/17538947.2020.1729879

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint.* , Article arXiv:1310.4546. https://doi.org/10.48550/arXiv.1310.4546

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint.* , Article arXiv:1511.08458. https://doi.org/10.48550/arXiv.1511.08458

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). https://doi.org/10.18653/v1/N18-1202

Qi, H., & Altinakar, M. S. (2011). Simulation-based decision support system for flood damage assessment under uncertainty using remote sensing and census block information. *Natural Hazards, 59*(2), 1125–1143. https://doi.org/10.1007/s11069-011-9822-8

Rhodan, M. (2017). Hurricane Harvey: The U.S.'s First Social Media Storm | Time. https://time.com/4921961/hurricane-harvey-twitter-facebook-social-media/.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint.* , Article arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47. https://doi.org/10.1145/505282.505283

Sit, M. A., Koylu, C., & Demir, I. (2019). Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane Irma. *International Journal of Digital Earth, 12*(11), 1205–1229. https://doi.org/10.1080/17538947.2018.1563219

Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 455–465). https://aclanthology.org/P13-1045.

Tang, L., & Liu, H. (2011). Leveraging social media networks for classification. *Data Mining and Knowledge Discovery, 23*(3), 447–478. https://doi.org/10.1007/s10618-010-0210-x

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Tellex, S., Katz, B., Lin, J., Fernandes, A., & Marton, G. (2003). Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 41–47). https://doi.org/10.1145/860435.860445

Turian, J., Ratinov, L.-A., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 384–394). https://aclanthology.org/P10-1040.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010).

Wang, C.-K., Singh, O., Tang, Z.-L., & Dai, H.-J. (2017). Using a recurrent neural network model for classification of tweets conveyed influenza-related information. In *Proceedings of the International Workshop on Digital Disease Detection Using Social Media 2017 (DDDSM-2017)* (pp. 33–38). https://aclanthology.org/W17-5805.

Wang, J., Hu, Y., & Joseph, K. (2020). NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS, 24*(3), 719–735. https://doi.org/10.1111/tgis.12627

Wang, K., Lam, N. S. N., Zou, L., & Mihunov, V. (2021). Twitter use in hurricane Isaac and its implications for disaster resilience. *ISPRS International Journal of Geo-Information, 10*(3), 116. https://doi.org/10.3390/ijgi10030116

Wang, S., Su, L., Li, S., Hu, S., Amin, T., Wang, H., Yao, S., Kaplan, L., & Abdelzaher, T. (2015). Scalable social sensing of interdependent phenomena. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks* (pp. 202–213). https://doi.org/10.1145/2737095.2737114

Wang, Z., Ye, X., & Tsou, M.-H. (2016). Spatial, temporal, and content analysis of twitter for wildfire hazards. *Natural Hazards, 83*(1), 523–540. https://doi.org/10.1007/s11069-016-2329-6

Xing, Z., Pei, J., & Keogh, E. (2010). A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter, 12*(1), 40–48. https://doi.org/10.1145/1882471.1882478

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). XLNet: generalized autoregressive pretraining for language understanding. *arXiv preprint.* , Article arXiv:1906.08237. https://doi.org/10.48550/arXiv.1906.08237

Yang, Z., Nguyen, L. H., Stuve, J., Cao, G., & Jin, F. (2017). Harvey flooding rescue in social media. *IEEE International Conference on Big Data (Big Data), 2017*, 2177–2185. https://doi.org/10.1109/BigData.2017.8258166

Yu, M., Huang, Q., Qin, H., Scheele, C., & Yang, C. (2019). Deep learning for real-time social media text classification for situation awareness – using Hurricanes Sandy, Harvey, and Irma as case studies. *International Journal of Digital Earth, 12*(11), 1230–1247. https://doi.org/10.1080/17538947.2019.1574316

Zhang, C., Lei, D., Yuan, Q., Zhuang, H., Kaplan, L., Wang, S., & Han, J. (2018). GeoBurst +: effective and real-time local event detection in geo-tagged tweet streams. *ACM*

Transactions on Intelligent Systems and Technology, 9(3), 34:1–34:24. https://doi.org/10.1145/3066166

Zou, L., Lam, N. S. N., Cai, H., & Qiang, Y. (2018). Mining Twitter data for improved understanding of disaster resilience. *Annals of the American Association of Geographers, 108*(5), 1422–1441. https://doi.org/10.1080/24694452.2017.1421897

Zou, L., Lam, N. S. N., Shams, S., Cai, H., Meyer, M. A., Yang, S., … Reams, M. A. (2019). Social and geographical disparities in Twitter use during Hurricane Harvey. *International Journal of Digital Earth, 12*(11), 1300–1318. https://doi.org/10.1080/17538947.2018.1545878