SN Computer Science (2022) 3:92
https://doi.org/10.1007/s42979-021-00971-4

ORIGINAL RESEARCH

Multi-source Multimodal Data and Deep Learning for Disaster Response: A Systematic Review

Nilani Algiriyage[1] · Raj Prasanna[1] · Kristin Stock[2] · Emma E. H. Doyle[1] · David Johnston[1]

Received: 28 July 2021 / Accepted: 11 November 2021 / Published online: 27 November 2021

Abstract

Mechanisms for sharing information in a disaster situation have drastically changed due to new technological innovations throughout the world. The use of social media applications and collaborative technologies for information sharing have become increasingly popular. With these advancements, the amount of data collected increases daily in different modalities, such as text, audio, video, and images. However, to date, practical Disaster Response (DR) activities are mostly depended on textual information, such as situation reports and email content, and the benefit of other media is often not realised. Deep Learning (DL) algorithms have recently demonstrated promising results in extracting knowledge from multiple modalities of data, but the use of DL approaches for DR tasks has thus far mostly been pursued in an academic context. This paper conducts a systematic review of 83 articles to identify the successes, current and future challenges, and opportunities in using DL for DR tasks. Our analysis is centred around the components of learning, a set of aspects that govern the application of Machine learning (ML) for a given problem domain. A flowchart and guidance for future research are developed as an outcome of the analysis to ensure the benefits of DL for DR activities are utilized.

Keywords Deep learning · Disaster management · Disaster response · Literature review

Introduction

Disasters, whether natural or human-induced, often result in loss of lives, property, or damage that can impose a significant impact on communities over a long period. With the proliferation of smart mobile devices, people are now increasingly using social media applications during disasters to share updates, check on loved ones, or inform authorities of issues that need to be addressed (e.g., damaged infrastructure, stranded livestock). Besides physical sensors and many other sources; human sensors, such as people who use smart mobile devices, generate massive amounts of data in different modalities (such as text, audio, video, and images) during a crisis. Such datasets are generally characterised as *multimodal* [17].

Disaster response (DR) tasks bring together groups of officials who often serve different organizations and represent different positions, and their information requirements remain complex, dynamic, and ad hoc [101]. Also, it is beyond the capacity of the individual human brain to combine different forms of data in real time and process them to form meaningful information in a complex and fast-moving situation [102]. Therefore, the main challenge faced by emergency responders is effectively extracting, analyzing, and interpreting the enormous range of multimodal data that is available from different sources within a short time period. As a result, emergency responders still depend mostly on text-based reports prepared by field officers for their decision-making processes, avoiding many other sources that could provide them with useful information.

✉ Nilani Algiriyage
r.nilani@massey.ac.nz

Raj Prasanna
r.prasanna@massey.ac.nz

Kristin Stock
k.stock@massey.ac.nz

Emma E. H. Doyle
e.e.hudson-doyle@massey.ac.nz

David Johnston
d.m.johnston@massey.ac.nz

[1] Joint Centre for Disaster Research, Massey University, Wellington, New Zealand

[2] Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand

Previously, the DR research community applied classical Machine Learning (ML) techniques to automate DR activities [2, 94]. However, the requirement of ML algorithms for handcrafted features prevented the timely use of such models. Furthermore, the research processes with these methods were labour-intensive and time-consuming [86]. More recently, Deep Neural Networks, which rely less on handcrafted features, instead learning directly from input data, have been used extensively to learn high-level representations through deep features and have proven to be highly effective in many application areas, such as speech recognition, image captioning, and emotion recognition [14, 17, 66, 119]. As DL techniques gain popularity among researchers, there is a timely need to discuss the potential for their use for DR activities. Researchers and practitioners need to understand what has been done in the literature and the current knowledge gaps to make further improvements. Thus, this article analyses and systematically reviews the intersection of the two research fields (DL for DR).

We have organized our review around the *components of learning* as proposed by Abu-Mostafa [121] and used by Watson et al. [125] for their systematic review. Abu-Mostafa [121] demonstrated the application of five components of learning for any ML problem. These components provide a clear mapping to establish a roadmap for investigating DL approaches in DR research. Our objective is to identify application scenarios, best practices and future research directions in using DL to support DR activities. Therefore, we synthesize five main Research Questions (RQs) and eight sub-questions that support the main RQs according to the components of learning. To answer the RQs, we create a data extraction form having 15 attributes such as DR Task, Data Type, Data Source, and DL Architecture. We create a taxonomy of DR tasks in response to the first RQ, which is then utilized to derive answers for the next RQs. Finally, we use the Knowledge Discovery in Databases (KDD) process to uncover hidden relationships among extracted values for the attributes in the extraction form. Based on our findings, we propose a flowchart with guidelines for DR researchers to follow when using DL models in future research.

We found multiple review articles that discussed the use of multimodal data for disaster response (for example, [6, 105]), outlining applications and challenges. However, many of these have not explicitly considered using DL for feature extraction. We also observed other review articles focused on individual data sources. For example, the studies [11, 55, 72, 91, 111, 124] addressed the frameworks, methodologies, technologies, future trends, and applications for disaster response while using social media datasets. Among other reviews, Gomez et al. [37] analyzed remotely sensed UAV data, considering cases of different disaster scenarios. Overall, these reviews are especially focused on addressing a single source of data and how it can be used for disaster response. The more recent article by Sun et al. [118] provides an overview of using Artificial Intelligence (AI) methods for disaster management. Our work significantly differs from the work by Sun et al. in a number of ways. Firstly, we analyze the articles systematically, adopting the learning components as proposed by Abu Moftha [121]. Secondly, our analysis is confined to trending DL techniques as a subset of AI. Thirdly, we provide a wider discussion on the datasets, preprocessing, DL architectures, hyperparameter tuning, challenges and solutions in processing data for the DL task, and clarify future research directions.

The remainder of this article is organized as follows. We first provide a synthesis of the research questions in Section "Research Question Synthesis". Section "Methodology" outlines the methodology used to analyze the literature. Sections "RQ_1: What types of DR Problems have been Addressed by DL Approaches?"– "RQ_5: What are the Underlying Challenges and Replicability of DL for DR Studies?" provide the analysis of the research questions and Section "Opportunities, Directions and Future Research Challenges" summarises opportunities and future research challenges. Section "Results of the Association Rule Mining" discusses the relationships extracted during the KDD process. In Section "Flowchart and Guidelines for Applying DL in Future DR Research" a flow chart is provided with recommendations for future research. Finally, in Section "Conclusions", we broadly discuss research gaps and conclusions. An online appendix contains the full details of the analysis process, as well as the resources [12].

Research Question Synthesis

Our overarching objectives during this study are to identify research challenges and best practices, and provide directions for future research while using DL methods for DR tasks. Therefore, we have centralized our analysis around the elements of learning (see Fig. 1) and formulated the main RQs accordingly. As a result, we ensure that our analysis effectively captures the essential components of DL applications while also allowing us to perform a descriptive content analysis across these components. Furthermore, we formulated sub-questions supporting the main RQs to analyze more details. The next subsections discuss the formulation of the main RQs and sub-questions according to the components of learning.

The First Component of Learning: The Target Function

The first component of the learning problem is an "unknown target function ($f : x \rightarrow y$)" as illustrated in Fig. 1, which represents the relationship between known input (x) and

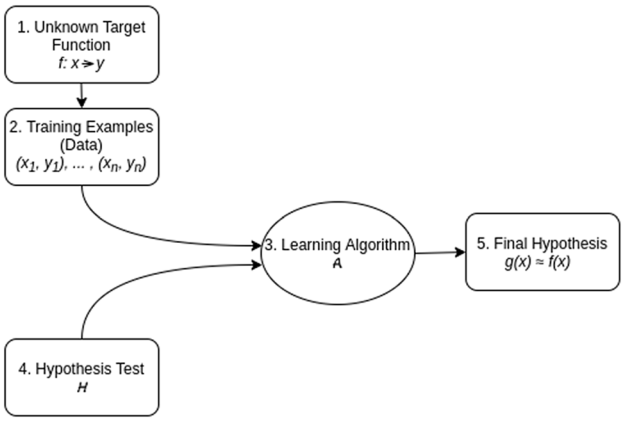


Fig. 1 The components of learning as proposed by Abu Moftha [121]

output (y). The Target Function is the optimal function that we are attempting to approximate with our learning model. Therefore, the first component of learning enables the researcher to identify main application areas in the research field. As a result, we formulated our first research question to identify target functions in the DR domain, as follows:

Our goal during this question is to capture the types of training data, the extraction sources, and the preprocessing techniques applied to prepare them for the DL tasks. To support and allow a deeper understanding of the main RQ, we examine this through three sub-questions.

- $RQ_{2.1}$ What types of DR data have been used?
- $RQ_{2.2}$ What sources have been used to extract data, and how have data been extracted?
- $RQ_{2.3}$ How have data been preprocessed before applying the DL models?

The answers we extract during questions $RQ_{2.1}$ and $RQ_{2.2}$ will enable future researchers to see what types and sources of data have been explored in previous studies and what data has not yet been investigated. Furthermore, $RQ_{2.3}$ provides them with the details of prepossessing techniques that have been followed during the studies.

RQ_1: What types of DR problems have been addressed by DL approaches?

RQ_1 aims to discover DR tasks that have been investigated previously using DL methodologies. Furthermore, the answers to our first RQ provide a taxonomy for analyzing the next research questions.

The Second Component of Learning: The Training Data

The second component of learning is the historical data (training data), required by the algorithm to learn the unknown target function. A thorough understanding of the training data leads to insights about the target function, possible features, and DL architecture design. Furthermore, the quality of the output of a DL model is directly coupled with the provided training data. Therefore, our second question is formulated to understand training data.

The Third and Fourth Components of Learning: The Learning Algorithm and Hypothesis Test

According to Abu Moftha [121], the third and fourth learning elements are known as the "learning model". The learning model consists of the learning algorithm and the hypothesis set. A learning algorithm tries to define a model to fit a given dataset. For example, the algorithm generally uses a probability distribution over the input data to approximate the optimal hypothesis from the hypothesis set. The hypothesis set consists of all the hypotheses to which the input data are mapped. Therefore, the learning algorithm and the hypothesis set are tightly coupled. Considering together the learning algorithm and hypothesis set, we formulate our third RQ as follows.

RQ_2: How have the training datasets been extracted, preprocessed and used in DL-based approaches for DR tasks?

RQ$_3$: What DL models are used to support DR tasks?

We aim to identify and evaluate the various DL models that have been applied for DR tasks. Hence, we consider three further sub-questions to capture specific architectures and types of DL models.

- $RQ_{3.1}$ What types of DL architectures are used?
- $RQ_{3.2}$ What types of learning algorithms and training processes are used?
- $RQ_{3.3}$ What methods are used to avoid overfitting and underfitting?

The answers to $RQ_{3.1}$ provide DL architectures that has been adopted for various DR tasks. Our goal is to determine whether certain DL architectures are preferred by researchers and the reasons for those trends. As a part of the analysis, we capture how transfer learning approaches have been adopted to address algorithm training and performance issues. During $RQ_{3.2}$, we intend to examine the types of learning algorithms and the training processes involved, including how parameter optimization has been achieved. Moreover, in $RQ_{3.3}$, we aim to analyze the methods used to combat overfitting and underfitting. Answers to both $RQ_{3.2}$ and $RQ_{3.3}$ will provide future researchers with an idea of how parameter tuning and optimization has been applied in DL for DR research to improve the accuracy of the output.

The Fifth Component of Learning: The Final Hypothesis

The final component of learning is the "final hypothesis". This is the target function learnt by the algorithm to predict unseen data points. Through this component of learning, we aim to analyze the effectiveness of the algorithm at achieving the hypothesis for the selected DR task. Therefore, our fourth RQ is formulated as follows:

During the analysis for **RQ$_4$**, we derive the metrics used to evaluate the performance of DL models. Future researchers can utilize these matrices and extract values to compare the results achieved by their models. Additionally, we examine two sub-questions to perform a deeper evaluation of the selected question.

- $RQ_{4.1}$ What evaluation matrices are used to evaluate the performance of DL models?
- $RQ_{4.2}$ What "baseline" models have been compared?

Our intention with $RQ_{4.1}$ is to derive a taxonomy of performance matrices used by the analyzed studies, while $RQ_{4.2}$ will identify those "baseline" models that have been criticized and allow future researchers to select those appropriate for comparison of their results.

The Final Analysis

Our fifth RQ is designed to identify and characterize underlying problems that arise when utilizing DL models for DR tasks. Our goal is to provide researchers with challenges faced by the DR research community in employing DL-based approaches. This will enable future research to be designed in a way that addresses or avoids these challenges and better utilizes DL algorithms to support DR tasks. Furthermore, we aim to analyze the replicability of DL models and architectures. Researchers are more likely to re-implement, improve, or compare new models if the existing DL architectures are easily replicable, which will eventually increase the quality and quantity of DL for DR research. Thus, our final RQ is formulated as follows:

RQ$_4$: How well do DL approaches perform in supporting various DR tasks?

RQ$_5$: What are the underlying challenges and the replicability of DL for DR studies?

In summary, the Systematic Literature Review (SLR) conducted in this paper answers the following research questions:

- **RQ_1:What types of DR problems have been addressed by DL approaches?**
- **RQ_2:How have the training datasets been extracted, preprocessed, and used in DL-based approaches for DR tasks?**
  - $RQ_{2.1}$ *What types of DR data have been used?*
  - $RQ_{2.2}$ *What sources have been used to extract data, and how have data been extracted?*
  - $RQ_{2.3}$ *How have data been preprocessed before applying the DL models?*
- **RQ_3: What DL models are used to support DR tasks?**
  - $RQ_{3.1}$ *What types of DL architectures are used?*
  - $RQ_{3.2}$ *What types of learning algorithms and training processes are used?*
  - $RQ_{3.3}$ *What methods are used to avoid overfitting and underfitting?*
- **RQ_4: How well do DL approaches perform in supporting various DR tasks?**
  - $RQ_{4.1}$ *What evaluation matrices are used to evaluate the performance of DL models?*
  - $RQ_{4.2}$ *What "baseline" models have been compared?*
- **RQ_5: What are the underlying challenges and the replicability of DL for DR studies?**

Methodology

Multiple techniques have been proposed to understand the content of a body of scholarly literature, including scoping reviews, umbrella reviews, or systematic reviews [38]. Among them, the systematic review aims to exhaustively and comprehensively search for research evidence on a topic area and appraise and synthesize it thoroughly [38]. In this analysis, we are interested in identifying the gaps in the research and whether there are opportunities for researchers and practitioners to investigate new problems that have not yet been addressed in the DR domain using DL. We, therefore, consider a systematic review to be the most appropriate approach to find answers to the above formulated RQs. To the best of our knowledge, this is the first systematic review that investigates the intersection of the DL and DR research fields. Our study adopts the following steps to guide the SLR process, as highlighted by Yigitcanlar et al. [128].

1. Develop a research plan.
2. Search for relevant articles.
3. Apply exclusion criteria.
4. Extract relevant data from the selected articles.
5. Analyse the literature data.

Develop a Research Plan

As the first step for carrying out the SLR, a research plan was developed, including research aim, keywords, and a set of inclusion and exclusion criteria. The research aim was to identify the usage of DL techniques on disaster data to support DR tasks as outlined in RQs 1–5. Hence, "disaster" and "deep learning" were selected as the search keywords. The search also included variants of these keywords. The alternate search terms for "disaster" included 'hazard', 'emergency', 'crisis', and 'damage'. Also, 'deep neural network' was used as an alternative keyword for DL. Some research has considered "machine learning" as an alternative keyword for DL. However, since we were particularly interested in Deep Neural Networks, we omitted "machine learning" as a keyword in the search. The inclusion criteria limited the sources to peer-reviewed academic publications available online in a full-text format and relevant to the research aims. The exclusion criteria were determined as publications in languages other than English; grey literature, such as government or industry reports; and non-academic research.

Search the Relevant Articles

In the second step, the search for relevant articles was conducted using a keyword search in each of the following databases: Scopus, Web of Science, and the EBSCO Discovery Service on April 2, 2021. Articles published since April 2011 were considered because a scan of existing literature suggested that there was not much literature related to DL in disaster research before then. The initial search produced 592 results.

Apply Exclusion Criteria

In this step, the results were filtered to remove duplicates between the databases, which reduced the number to 295 unique articles. We used a simple Python script to remove

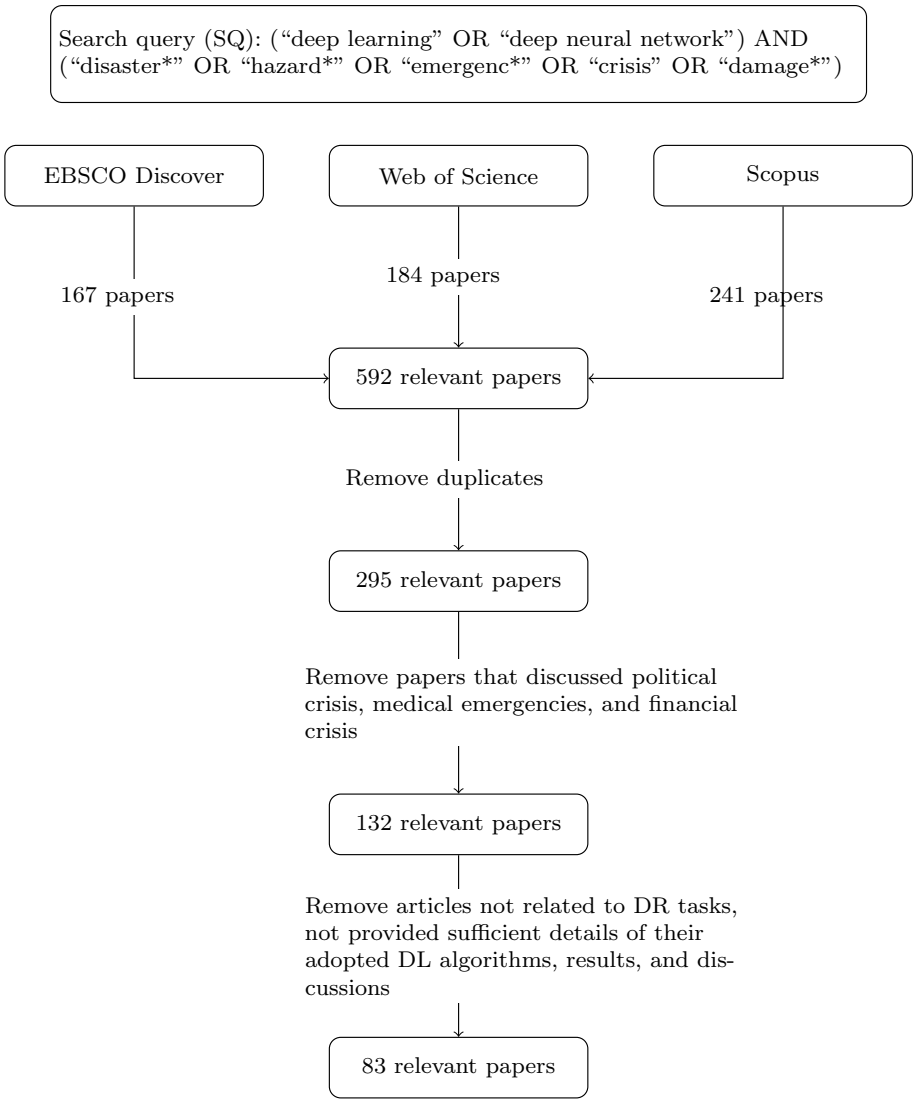


Fig. 2 Literature selection process

duplicates using the title of the article. We confined our scope to only papers that discuss natural or human-induced disasters. Therefore, the abstracts were manually read and removed if they discussed political crises, medical emergencies or financial crises. We also removed articles that did not provide sufficient details related to the attributes in our extraction form (see Table 1). Finally, 83 articles were selected for the review. Fig. 2 illustrates the process and the steps that we followed to filter the results and the quantity of papers returned at each step. Moreover, we provide the publication venues of the 83 articles in Fig. 3.

Extract Relevant Data from the Selected Articles

The next step in our methodology was to extract relevant data from the selected articles. We developed a data extraction form including the information shown in Table 1. The extracted information was collected manually and added to a Google sheet and later downloaded as a tab-separated (.tsv) file for the data analysis steps. The extracted data sheet is available in the online appendix [12].

Analyse Data Using the Knowledge Discovery in Databases (KDD) process

The final step in our SLR methodology was to analyze the extracted data. We used the steps discussed in [125], namely data collection, initial coding and focused coding. After the coding process, we used the Knowledge Discovery in Databases (KDD) process to understand relationships among attributes in the extraction form. The KDD process is used to extract knowledge from databases using five steps: selection, preprocessing, transformation, data mining, and interpretation/evaluation [33]. We combined data preprocessing and transformation into one step as both steps involve preparing

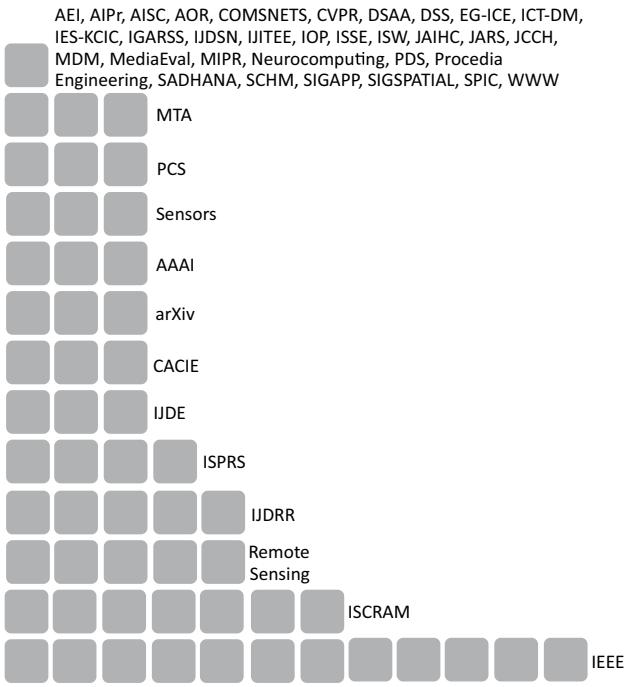


Fig. 3 Publication venues of the articles. The number of grey boxes corresponds to the number of articles published in each publication venue. Full publication venue names are available in the Appendix B

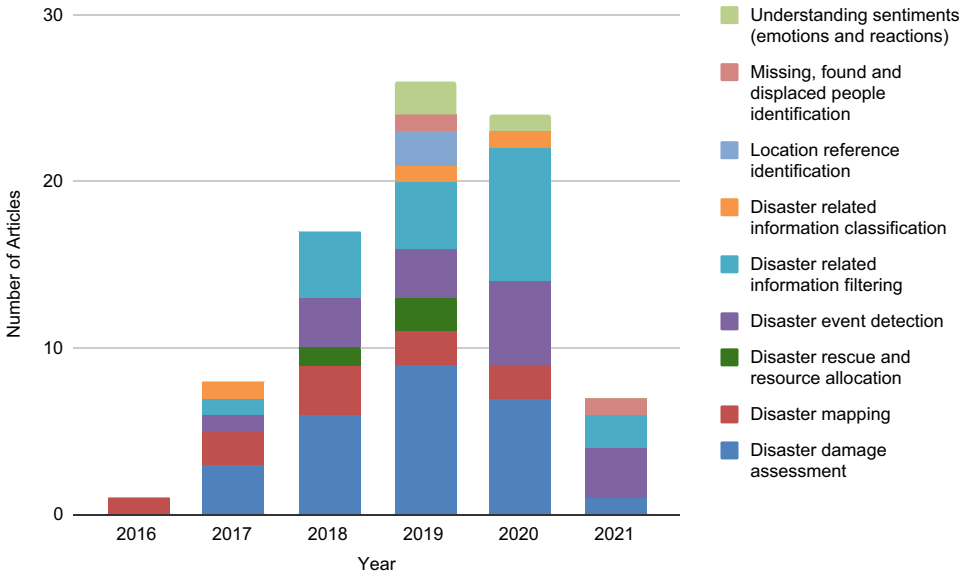
data for the mining step. The details of each stage are listed as follows.

- *Selection:* This stage is related to the selection of relevant data for the analysis. As described in the previous section, we selected 83 articles and extracted 15 attributes from them for the analysis.
- *Preprocessing:* In this stage, we cleaned the extracted values by removing noise, such as misspellings, incorrect punctuation and mismatching coding. We noticed that a number of variations on particular terms, and standardized these to ensure appropriate matching (e.g., ConvNet/CNN, *F*-measure/*F*1-value/*F*-score/*F*1-score).
- *Data mining:* The third stage is related to identifying relationships among extracted data. We applied association rule mining to derive relationships discussed further in Section "Association Rule Mining".
- *Interpretation/Evaluation:* We interpret the findings of the KDD process in Section "Results of the Association Rule Mining". These relationships demonstrate actionable knowledge for future researchers from the 83 articles analyzed through the SLR process.

Table 1 Attributes in the data extraction form

Article published year	Venue	DR task addressed
Input data modality	Data source	Data extraction Technique
Data preprocessing technique	Size of the dataset	Type of learning
DL architecture used	Learning algorithm	Evaluation metrics
Replicability	Baseline	Combating overfitting and underfitting

Fig. 4 Papers published per year according to DR task



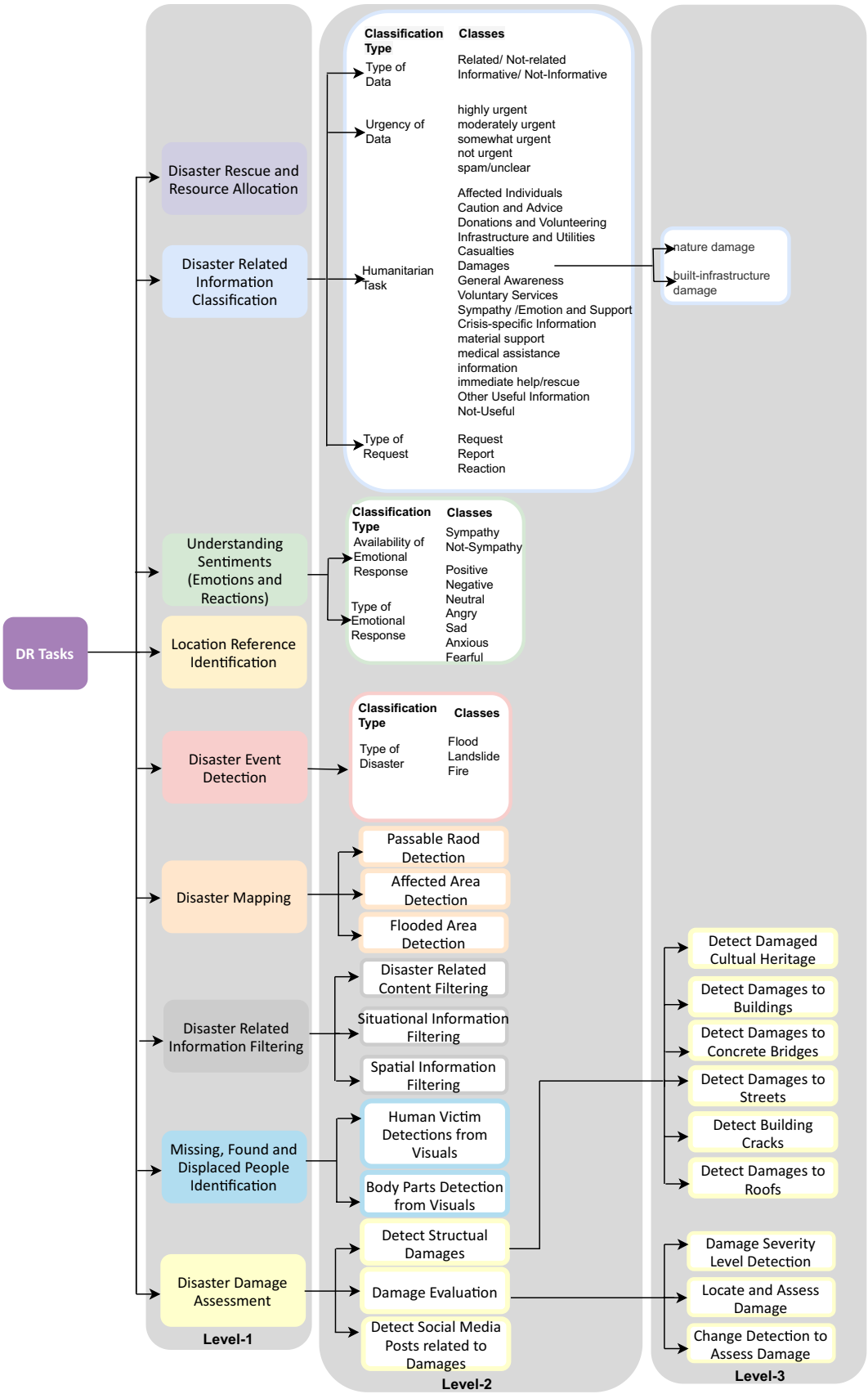


Fig. 5 Taxonomy of DR Tasks

Table 2 Main DR tasks of the analysed articles

DR task	Articles
Disaster related information filtering	[8, 21, 34, 43, 45, 46, 60, 63, 73, 75–77, 81, 87, 93, 95–97, 115]
Disaster damage assessment	[3, 16, 24, 26, 27, 31, 35, 36, 39, 44, 51, 59, 62, 64, 67, 68, 78, 79, 82, 88, 89, 103, 112, 117, 122, 130]
Disaster event detection	[7, 13, 15, 53, 69–71, 83, 90, 129]
Location reference identification	[56, 113]
Missing, found and displaced people identification	[41, 80]
Disaster mapping	[4, 85, 106]
Disaster rescue and resource allocation	[19, 20, 30]
Understanding sentiments (emotions and reactions)	[65, 110, 127]
Disaster related information classification	[1, 5, 9, 22, 25, 52, 57, 58, 74, 92, 99, 100, 108, 109, 120]

Association Rule Mining

We followed the association rule mining process introduced by Samia et al. [61] for literature analysis. Our association rules are extracted using the Apriori algorithm. Association rules help to discover relationships in categorical datasets. For instance, the rules generated during the process identify frequent patterns in the dataset. Associations are generally represented by "Support", "Confidence", and "Lift". We illustrate this using the values in the *Data Source* column in the extraction form. "Support" and "Confidence" are the two indicators evaluating the interestingness of a given rule. *Supp(Twitter)* is the fraction of articles for which **Twitter** appears in the *Data Source* column of the extraction form as given in Eq. 1.

$$supp(Twitter) = \frac{\text{Number of Articles in which } Twitter \text{ appears in the Data Source column}}{\text{Total Number of Articles}}. \quad (1)$$

If we consider the values in both the *Data Source* and the *Data Type* columns of the extraction form, the association rule *Twitter*→ *Text* means that each time **Twitter** appears in the *Data Source* column, **Text** appears in the *Data Type* column (see Eq. 2).

$$conf(Twitter \rightarrow Text) = \frac{supp(Twitter \cup Text)}{supp(Twitter)}. \quad (2)$$

"Lift" measures how likely it is that item **Text** is found in the *Data Type* column when **Twitter** is found in the *Data Source* column as given in Eq. 3. A "Lift" value greater than 1 means that item **Twitter** is likely to appear in the *Data Source* column if **Text** appears in the *Data Type* column, while a value less than 1 means that **Twitter** is unlikely to appear if **Text** appears in the respective columns.

$$lift(Twitter \rightarrow Text) = \frac{supp(Twitter \cup Text)}{supp(Twitter) \times supp(Text)}. \quad (3)$$

These associations can provide a guidance for future researchers during the planning stages of a project applying DL to DR research, supporting them in choosing different attributes, such as data source, deep learning algorithm and learning types. We used the Python apyori library[1] to discover association rules, details of which are presented in the online appendix [12].

RQ_1: What types of DR Problems have been addressed by DL Approaches?

This RQ explores the types of DR problems that have been investigated with DL models. We derived a taxonomy of DR tasks to capture relationships between other learning components, as illustrated in Fig. 5. From the 83 papers that we analysed, we identified nine main DR tasks (level-1 of the taxonomy) that have been addressed using DL approaches. Figure 4 shows the number of papers published in each year by the main DR tasks. During the ten-year duration of studies we analysed, unsurprisingly, little work was undertaken between 2011 and 2015. There was a sudden interest in exploring DL architectures in the DR domain from 2017 onwards. This interest coincides with the introduction of popular DL frameworks, such as Keras[2] and TensorFlow[3] in 2015 and PyTorch[4] in 2016. *Disaster event detection* was the first task to be explored using DL algorithms. Among the other tasks, *Disaster damage assessment, Disaster-related*

[1] Python Apriori algorithm implementation v1.1.2, https://pypi.org/project/apyori/.

[2] Keras, https://keras.io/.

[3] TensorFlow, https://www.tensorflow.org/.

[4] PyTorch, https://pytorch.org/.

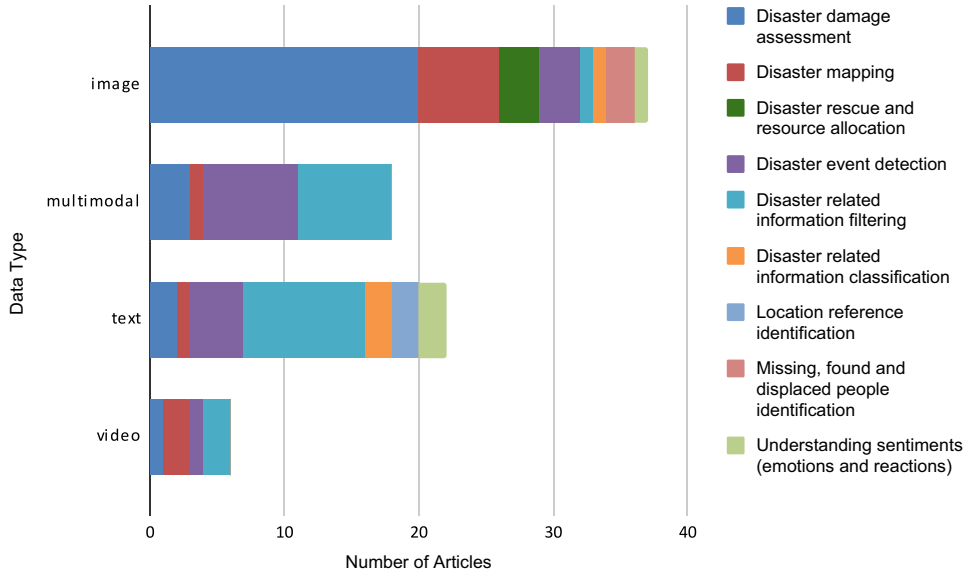


Fig. 6 Data types used for DR task

information filtering and *Disaster-related information classification* were explored in 2017. Remotely sensed images were the main source of data for multiple early studies that used DL approaches. Early research may have used remotely sensed data for various reasons. Firstly, in 2011, Google Earth[5] launched a platform that allowed researchers to download massive volumes of satellite imagery. This inspired researchers to investigate remotely sensed data for DR tasks. Furthermore, researchers were also able to successfully employ DL approaches since these images were available in larger quantities. Secondly, the advancement of computer vision techniques, such as DL structures pretrained on huge datasets, made visual data processing easier.

The number of studies combining DL and DR tasks rapidly increased from 2017 to 2018, more than doubling. Furthermore, researchers extended their interest to explore multiple DR tasks over time, including *Disaster rescue and resource allocation*, *Location reference identification*, and *Understanding sentiments*. However, we see a slight drop in the number of articles published in 2020. This inconsistency may be due to the COVID-19 global pandemic and the physical and mental challenges that researchers encountered. We notice a significant amount of literature emerging during the first quarter of 2021, potentially representing a COVID-19 lag effect in publication.

Disaster damage assessment has been the most popular DR task analysed using DL approaches over the years, with 26 articles out of the 83 exploring this. There are three likely reasons for the popularity of *Disaster damage assessment*. First, there is quite a strong driver and a clear need for damage assessment as it is urgently needed following an event, and there is little time for manual data collection. Second, the high availability of training datasets extracted from social media and remote sensing platforms was able to be used in supervised learning approaches. Third, there is a clear mapping between training data and the target function (e.g., images of cracked buildings). This mapping helps researchers when designing DL-based applications to extract effective features. We observed an increasing interest in *Disaster-related information filtering* and *Disaster-related information classification* tasks. These DR tasks are mainly based on text datasets extracted from Twitter. A possible explanation for this trend could be the increased popularity of using Twitter as a communication channel during disasters. Moreover, the advancement of Natural Language Processing (NLP) techniques with the increased availability of annotated data corpora aids further developments in the information filtering and classification tasks.

DR tasks, such as *Missing, found and displaced people identification* and *Location reference identification*, had received less attention from researchers, resulting in a total of 4 articles out of the 83 reviewed. The lack of availability of large-scale training datasets and annotated data to train supervised learning approaches could be the main reasons for the reduced popularity of these DR tasks. We summarise the papers addressing each of the main DR tasks in Table 2.

RQ_2: How have the Training Datasets been Extracted, Preprocessed, and Used in DL-Based Approaches for DR Tasks?

For this research question, we analyze the types of disaster data that have been used by DL models to support disaster response. The accuracy and effectiveness of DL algorithms

[5] Google Earth, https://earth.google.com/web/.

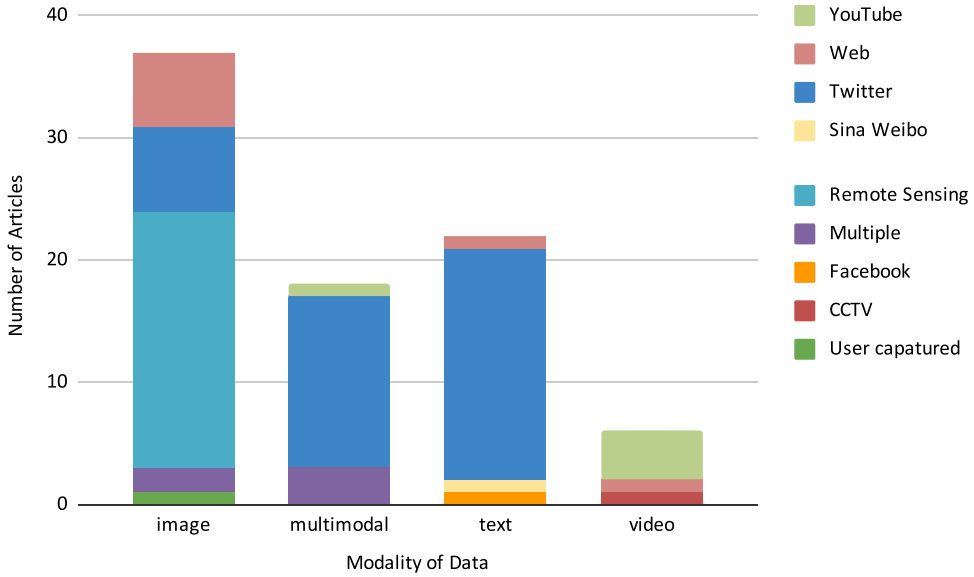


Fig. 7 Sources used to extract data types

depend on the training dataset and its clarity. Therefore, we aim to understand the various types of disaster data used by DL approaches, the sources and methods employed to extract them, and the preprocessing steps. All of these points are important in understanding and designing DL approaches for DR tasks.

$RQ_{2.1}$ What Types of DR Data have been used?

Our analysis of the types of data that have been used for DR tasks using DL approaches reveals relationships between DR tasks and data types, illustrated in Fig. 6. Among the 83 articles analysed, 37 used images as the data source. Surprisingly, in practise, disaster responders rely significantly on textual data sources, such as emails and field reports [39]. This finding indicates that these approaches have been mostly pursued in academic contexts. We assume multiple reasons contributing to the popularity of using image data for DR tasks: first, the power of visuals in conveying messages over textual content; second, the availability of pretrained networks and the use of transfer learning techniques for image feature extraction and third, easy accessibility of image datasets through web search and web databases. *Disaster damage assessment* is the most popular DR task among the studies that used image datasets.

Text data were used by 22 of the 83 articles and is more prominent in *Disaster-related information filtering* and *Disaster-related information classification* tasks. Currently available, annotated disaster-related text data repositories (particularly using social media data) provide a clear guide for specific target problems. As a result, many researchers have used text data for supervised learning approaches in information filtering and classification applications.

There has been little interest in using video datasets for DR tasks. Only 6 articles discussed the usage of video datasets for *Disaster related information filtering*, *Classification*, and *Disaster event detection* tasks. The possible reasons for this can be difficulties in storing and moving, and the need for special computing facilities for analysing video data such as Graphical Processing Units (GPUs).

We observed a significant interest in using multimodal data to extract information for DR tasks between 2018 and 2020. Multimodal data have been used for *Disaster-related information filtering*, *Disaster-related information classification*, *Disaster damage assessment* and *Disaster event detection* contributing to 18 articles in the analysed papers. We assume the popularity of multimodal DL networks depends on three reasons. First, the combination of multiple modalities leads to more complementary information than learning from a single data modality. Second, multimodal learning helps to integrate data from different sources and provides access to large quantities of data. Third, the more recent development of multimodal DL networks shows improved results over unimodal analysis.

$RQ_{2.2}$ What Sources have been Used to Extract Data, and How have Data been Extracted?

In this RQ, we analyse the sources (including accessible disaster data repositories) used to extract data used in DL models.

Image data have mainly been extracted using remote sensing from sources, such as satellites, aerial vehicles and LiDAR. Apart from that, Twitter and the Web have been used by 7 and 6 articles, respectively, to extract image datasets (we grouped research that extracted data from websites and Google search under *Web*). Twitter has been the

Table 3 Disaster data collection methods

Data extraction method	Articles
Artificial Intelligence for Disaster Response (AIDR)	[9, 89]
Baidu API	[127]
Cameras mounted on satellite, airborne and UAV	[16, 20, 24, 31, 67, 88, 106, 122]
Copernicus EMS program	[51]
CrisisLex	[21, 22, 82, 87, 90]
CrisisMMD	[1, 3, 34, 57, 58, 63, 74–76, 100]
CrisisNLP	[5, 65, 82, 90]
Facebook page crawling	[97]
Flicker API	[39]
GNIP (Social media data re-seller)	[30, 113]
Google Earth	[13, 26, 69, 70, 85]
LiDAR	[112]
Previous research	[15, 62, 64, 68, 73, 130]
Twitter API	[9, 19, 39, 43, 52, 56, 77, 81, 93, 109, 110, 115]
Web database	[25, 35, 36, 41, 44, 53, 60, 79, 83, 108, 117]
Web mining	[8, 27, 45, 46, 59, 64, 71, 78, 80, 82, 89, 95, 96, 99, 120, 129]
Workshop/Conference	[4, 7, 18, 103]

prominent source of text information, and was used for a total of 19 articles out of the total 83 (and out of the 22 articles that used text data) analysed. The growing number of human-annotated disaster-related Twitter data repositories is likely to have increased the amount of research using them with DL approaches. We observed that 5 articles used a combination of multiple sources to extract data, such as Twitter, web mining, Baidu, Flicker, Instagram, and Facebook. Most notably, Facebook was rarely used (1/83) as a source due to its data extraction limitations (e.g., the requirement of prior approval from Facebook to use public feed Application Programming Interface (API) [104]). Figure 7 shows the sources used to extract different modalities of data.

Researchers have employed multiple techniques to extract data from different sources. Twitter data have been extracted through the Twitter Streaming API using general or specific keywords (e.g., earthquake, Nepal Earthquake) and a spatial bounding box covering the impacted area is often used while extracting tweets. However, it is notable that a total of 28 articles downloaded data from annotated Twitter repositories from previous research, such as CrisisNLP[6] and CrisisLex,[7] indicating the importance of annotated data repositories catering for DR problems. Web mining and web databases were used in 22 articles to download data. Workshops and conferences, for example, MediaEval,[8] have provided researchers with annotated dataests and meta-data for target problems. Table 3 summarizes the different data collection methods.

$RQ_{2.3}$ How have data been Preprocessed Before Applying the DL Models?

To address $RQ_{2.3}$, we derive a taxonomy of preprocessing steps that researchers have used to clean raw data for use in DL algorithms. Cleaning and transforming data to be used effectively by DL models are critical steps towards improved performance. However, 19 articles out of 83 analysed did not explicitly mention the preprocessing steps that were undertaken.

We observe three common preprocessing steps across the articles analyzed: filtering, annotation, and dataset splitting. Data filtering helps reduce noise in raw data. Annotation deals with labelling the data depending on the target function. A total of 10 of the 83 articles employed external annotators or hired them through annotation service providers such as Figure Eight[9] (formerly known as CrowdFlower). The annotated datasets are generally split into train, test, and validation sets during the preprocessing steps. The training data sets are used to train the DL model, while test datasets are used to provide unseen data to be classified by the model as a test. The validation set is used to tune hyperparameters of the DL model.

[6] CrisisNLP datasets, https://crisisnlp.qcri.org/.

[7] CrisisLex datasets, https://crisislex.org/.

[8] MediaEval datasets, http://www.multimediaeval.org/.

[9] Figure Eight external annotation service, https://appen.com/.

Table 4 Data preprocessing steps

Modality	Preprocessing step	Description/Example
Text	Tokenizing	Tokenization is the process of breaking sentences in to smaller chunks (e.g., words)
	Lowercasing	Lowercasing tweet text is used to merge similar words and reduce the dimensionality of the problem
	Removal of stop words	Stopwords are a set of frequently used words such as *"the"*, *"in"*, and *"a"* that are not required to analyse them for a analysis task
	Removal of URLs and user mentions	Tweets generally consist user handlers and embedded URLs. During preprocessing, they are removed or replaced with < *USER* > and < *URL* >respectively
	Removal of hashtags	Hashtags are words or phrases chosen by users to connect specific themes such as events and topics (e.g., *#NepalEq*)
	Removal of punctuation, whitespaces, linebreaks	Punctuations (e.g., ".!@#":;"), whitespaces and linebreaks are removed as they do not contain valuable information for a analysis task
	Removal of numbers	Numerical values included in tweets are removed if they do not contain any information for the analysis task
	Removal of words shorter than 3 characters	Shorter words such as *"oh"*, *"omg"* and *"hmm"* are not useful for the analysis task and therefore, removed
	Replacing contractions	The user-generated Twitter posts mostly contain shorten phrases (e.g., *I'd, didn't and I'll've*). During the contraction mapping, these words are mapped into their original format (e.g., *I would, did not, I will have*)
	Stemming and lematization	Stemming and lemmatization are used to convert a word into its root format. The stemming process cuts off the ends of words without considering the context, while lemmatization considers the context (e.g., *felt to feel*)
	Remove sentences having less than three words	Remove very short sentences
Image	Manual filtering	Manually check images to remove unwanted
	Patch generation	Select arbitrary shaped regions from an original image
	Resizing	
	Pixel value normalization	Pixel values of an image normally are between 0-255. During the normalization, values are converted to be in a specified range such as [1-0]
	Image transformation	(e.g., rotation, translation, rescaling, flipping, shearing, and stretching)
Video	Manual filtering	Manually check videos to remove unwanted
	Shot boundary detection	A shot is an unbroken sequence of frames and a shot boundary is determined by the change of color histogram features
	Clipping to extract key frames	Extract frames in the middle of each shot as key frames
	Removal noisy frames	Remove duplicates and blurred frame

Our analysis identified that the design of the preprocessing steps largely depends upon the modality of data. For example, text data preprocessing steps included tokenizing, lowercasing, stemming, lemmatization and removal of stop words, tokens having less than 3 characters, sentences having less than 3 words, user mentions, punctuation, extra spaces, line breaks, emojis, emoticons, special characters, symbols, hashtags, numbers, and duplicates. Text normalization using the Out of Vocabulary (OOV) dictionary is used to replace slang, mistakenly added words, abbreviations, and misspellings. Image data preparation steps included data filtering, duplicate removal, patch generation, resizing, pixel value normalization, and image augmentations. Video data preprocessing included clipping to extract keyframes, shot boundary detection and removal of duplicates and blurred and noisy frames. Table 4 illustrates the preprocessing steps involved in preparing raw data for DL algorithms, as found in the analyzed articles.

RQ_3: What DL Models are Used to Support DR Tasks?

In this section, we analyze the types of DL architectures used for DR tasks and learning algorithms. Our aim is to identify the relationship between DR tasks and the DL architectures. We provide a short overview of different deep learning architectures in our online appendix [12].

$RQ_{3.1}$ What Types of DL Architectures are Used?

Through this question, we analyze types of DL architectures used to extract features for DR tasks. We observed that six

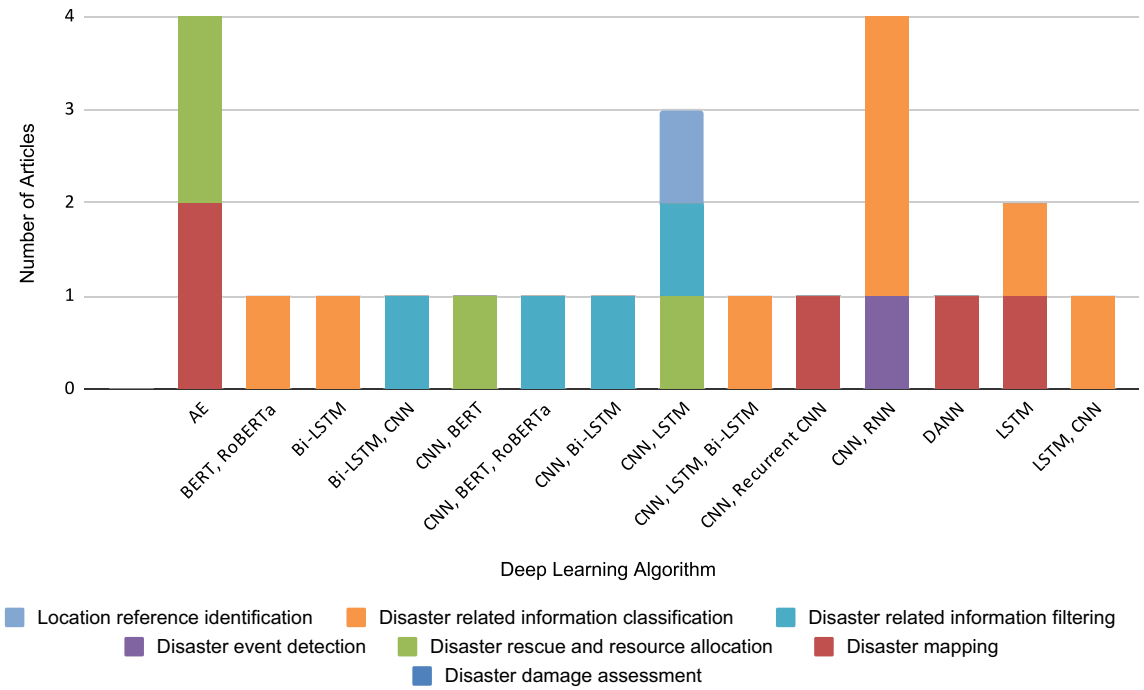


Fig. 8 DL architectures used by DR tasks except for CNN as a single architecture

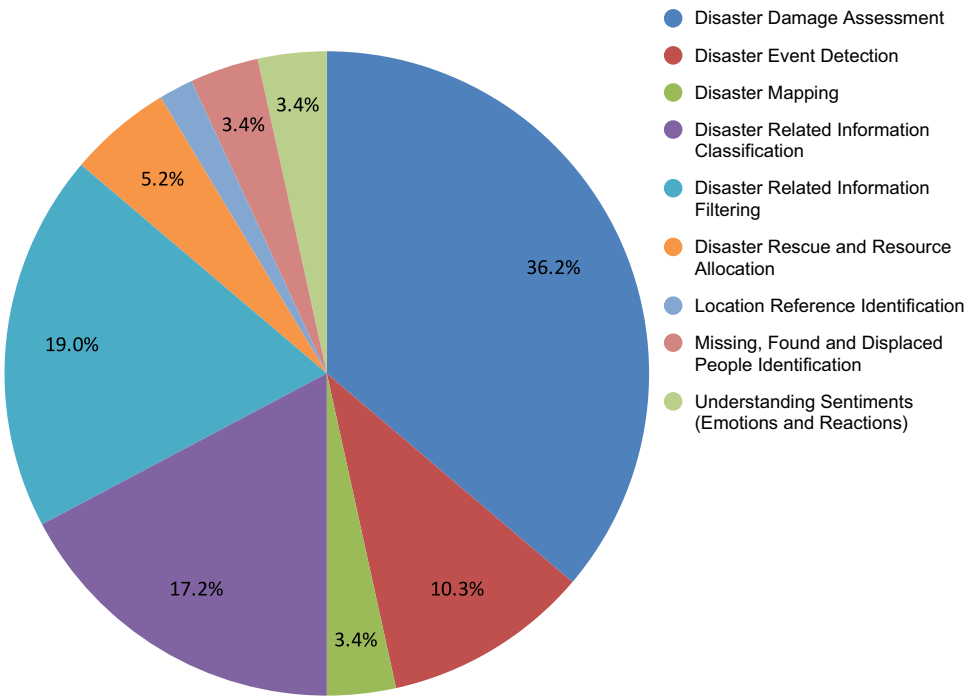


Fig. 9 Usage of CNN by DR tasks

main DL architectures had been used, namely Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs) and its variant Bi-directional LSTMs (Bi-LSTMs), Domain Adversarial Neural Networks (DANNs), and AutoEncoders (AEs) across the studies we analyzed. Moreover, popular language models like Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pre-training Approach (RoBERTa) have been used for Natural Language Processing (NLP) tasks.

Figure 8 shows the usage of DL algorithms according to the DR tasks excluding CNNs. We demonstrate the

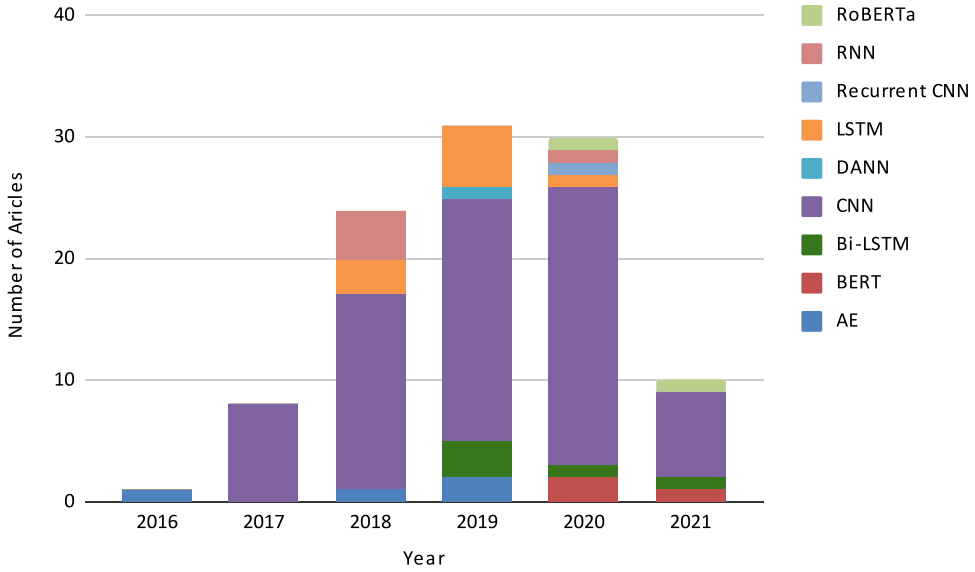


Fig. 10 DL architectures used by DR tasks by year

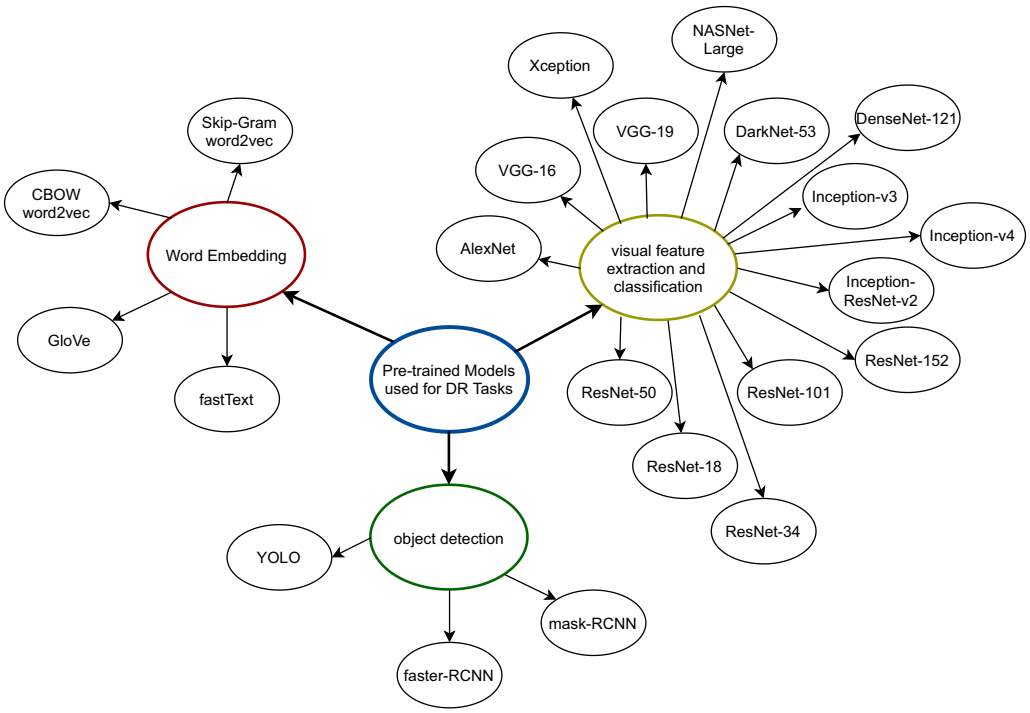


Fig. 11 Pre-trained DL networks used by DR tasks

application of the CNN algorithm for DR tasks in a separate diagram (see Fig. 9), and we present the usage of DL architectures based on publication year in Fig. 10. There has been a significant growing interest in using CNNs over the years across all DR tasks in 71 out of 83 articles analyzed. We consider it likely that CNNs have been adopted largely due to their capability in learning features automatically, parameter sharing and dimensionality reduction [114]. However, CNNs have performed poor for identifying word order in a sentence for text classification tasks [73]. Moreover, the computational cost (e.g., training time) for CNNs has been considerable, particularly when the training dataset is large.

RNNs, LSTMs, and Bi-LSTMs have been used to analyze varying length sequence data such as sentences (e.g., tweet text). Although RNNs have been successful in many sequence prediction tasks, it has issues in learning long-term dependencies due to the vanishing gradient problem. This problem occurs from the gradient propagation of the

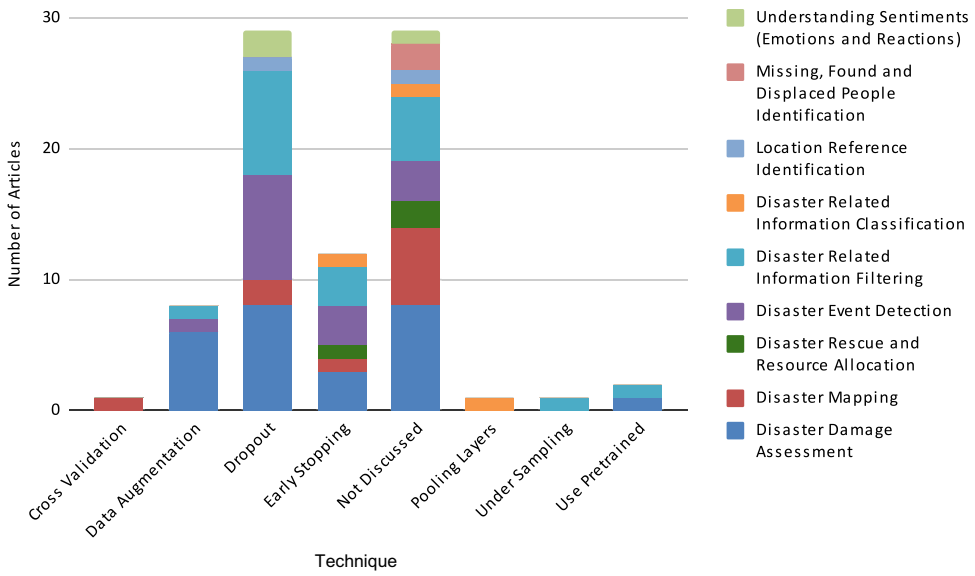


Fig. 12 Methods used to avoid overfitting and underfitting by DR tasks

recurrent network over many layers [73]. LSTM networks have been proposed to overcome these drawbacks and have shown better results for multiple text classification tasks [99]. Recent studies have demonstrated more improved results using Bi-LSTMs. One of the major advantages of using Bi-LSTMs is that they can capture and deal with long-range dependencies having variable lengths by analyzing information in both directions of a sequence (e.g., past and future entries) [43, 52].

We observe that many studies adopt DL models pre-trained on larger data sets, such as Places365[10] and ImageNet.[11] Fifty-one of the analyzed papers used pre-trained DL networks for word embeddings, visual feature extraction, object detection and classification. The advantage of adopting a pre-trained model is that it saves time and resources relative to training a model from scratch. Figure 11 provides a taxonomy of pre-trained networks adopted by our analyzed studies.

In addition, we observed that 17 studies adopted multiple DL architectures. This is very common in research that uses different modalities of data. For example, CNNs are often used to extract image features, while RNNs, LSTMs or Bi-LSTMs are used for text feature extraction.

$RQ_{3.2}$ What Training Processes are Used to Optimize DL Models?

In this RQ, we analyze the processes used to train DL algorithms focusing on optimization and error calculation.

All but four of the 83 articles used supervised learning as the training type for the selected DR problem. In supervised learning, the DL algorithm extracts features to associate data with the required classification labels. Therefore, a labelled training dataset is required. In contrast, unsupervised learning assigns a class label by grouping similar data together based on extracted features. Therefore, unsupervised approaches do not require labelled training data. Semi-supervised approaches use partially labelled data sets. However, both unsupervised and semi-supervised approaches were rarely used in the analyzed articles resulting in only 4/83. The current favour for supervised learning approaches is mostly due to the readily available labelled datasets. However, those outdated datasets would not reflect the temporal variations, and therefore, more improvements are required for DL architectures to make approximations without training.

The classical gradient descent algorithm was the most frequently adopted learning algorithm in the articles we analyzed for updating weights during backpropagation. Although researchers widely use gradient descent, the computational complexity is considerable because the entire dataset is considered every time the parameters are updated [98]. Multiple other algorithms, such as Adaptive Moment Estimation (Adam), Adadelta, and RMSProp algorithms, were proposed to overcome this issue. These new techniques have been used for optimization by 45 articles. The selection of optimization algorithm significantly affects the results of the model. However, we could observe that only 31% of the analyzed articles explicitly mention the optimization process and the algorithms they used.

Our analysis found that multiple algorithms have been adopted to calculate the error rate. Categorical cross-entropy is the most frequently used loss function, while negative log-likelihood was adopted by one article. The objective of a loss function is to optimize and tune weights in deep neural

[10] Places365 dataset, http://places2.csail.mit.edu/download.html.

[11] ImageNet dataset, https://image-net.org/.

network layers. However, only 22 of the papers discussed the error function.

$RQ_{3.3}$ What Methods are Used to Avoid Overfitting and Underfitting

Two common problems associated with generalizing a trained DL model are known as "overfitting" and "underfitting". Overfitting happens when the model learns training data extremely well but is not able to perform well on unseen data [42]. In contrast, an underfitted model fails to learn training data well and hence performs poorly on new unseen data. This happens due to the lack of capacity of the model or not having sufficient training iterations [49]. In both these cases, the model is not generalized well for the target problem.

To combat overfitting and underfitting, we observed that research had used multiple techniques such as *Dropout, Batch normalization, Early stopping, Pooling layers, Cross-validation, Undersampling, Pre-trained weights* and *Data augmentation*. Figure 12 illustrates these methods by DR tasks. A total of 24 articles used *Dropout* layers and 12 articles used *Early stopping* to avoid overfitting. Dropout layers ignore nodes in the hidden layer when training the neural network, and therefore, it prevents all neurons in a layer from optimizing their weights [116]. However, the batch normalization technique was proposed to achieve higher accuracy with fewer training steps, eliminating the need for Dropout [48]. During model training, the Early stopping technique evaluates the performance of the model on the validation dataset. The training process is stopped when the accuracy starts decreasing. As a result, however, this technique prevents the use of all available training data. Rice et al. [107] provide remedies for overfitting using a series of experimental evaluations.

Addressing underfitting while training DL models is a complex task, and these are not well-defined techniques [125]. We observed that 2 articles used pre-trained weights to avoid underfitting. However, 29 of the analyzed articles did not discuss the methods used for combating overfitting or underfitting.

RQ_4: How well do DL Approaches Perform in Supporting Various DR tasks?

In this RQ, we analyze the effectiveness of DL approaches for DR tasks, including reviewing the evaluation matrices and baseline models and comparing results achieved.

$RQ_{4.1}$ What Evaluation Matrices are Used to Evaluate the Performance of DL Models?

Through this question, we explore the different performance matrices adopted by the studies we analysed. Our aim is to identify how the existing research evaluated their results. Evaluation of the performance of a model is a core function when employing DL algorithms, as it helps to improve the model constructively. We observed that 76 of the 83 articles had adopted standard performance evaluation matrices, such as precision, recall, accuracy, and $F1$-score (see the definition of these metrics matrices in Eqs. 4– 9). These measures are based on the "true positive", "false positive", "true negative", and "false negative" values, which evaluate the correctness of the results.

	True condition		
Predicted condition	**True positive** T_p	**False positive** F_p	Precision/Positive Predictive Value (PPV) $\frac{T_p}{T_p+F_p} \times 100\%$
	False negative F_n	**True negative** T_n	Negative Predictive Value (NPV) $\frac{T_n}{T_n+F_n} \times 100\%$
	Sensitivity/Recall Rate (RR) $\frac{T_p}{T_p+F_n} \times 100\%$	Specificity Rate (SR) $\frac{T_n}{T_n+F_p} \times 100\%$	

$$\text{Precision/Positive Predictive Value (PPV)} = \frac{T_\text{p}}{T_\text{p}+F_\text{p}} \times 100\% \tag{4}$$

$$\text{Recall/Sensitivity} = \frac{T_\text{p}}{T_\text{p}+F_\text{n}} \times 100\% \tag{5}$$

Table 5 Best accuracy scores for DR tasks

Author	DR task	Sub-task	Best Accuracy Score			
			Precision	Recall	Accuracy	*F*1-score
[65]	Understanding Sentiments (Emotions and Reactions)	Classification-binary (e.g., Sympathy vs Non-Sympathy)	0.95		0.71	0.76
[110]		Classification-multiclass (e.g., Angry, sad, anxious, fearful)	0.90	0.88	0.93	0.89
[41]	Missing, Found and Displaced People Identification	Human Victim Detection from Visuals			1.00	
[80]		Body Parts Detection from Visuals	0.96	0.99	0.95	
[56]	Location Reference Identification		0.97	0.95	0.96	
[4]	Disaster Mapping	Passable Road Detection				0.65
[85]		Affected Area Detection			0.92	
[30]	Disaster Rescue and Resource Allocation		0.94	0.92	0.98	0.87
[30]	Disaster Event Detection	Flood Detection				0.86
[69]		Landslide Detection	0.98	0.97	0.97	
[129]		Early Fire Detection			1.00	
[24]	Disaster Damage Assessment	Structural Damage Detection	0.88	0.95	0.99	0.91
[39]		Damage Evaluation	0.85	0.78	0.99	
[3]		Damage-related Social Media Posts Detection	0.99	0.99		0.99
[74]	Disaster Related Information Classification	Classification-binary (e.g., Informative vs Not-Informative)			0.96	0.96
[5]		Classification-multiclass (e.g., Affected Individuals, casualties, damages)			0.97	
[60]	Disaster Related Information Filtering	Disaster Related Content Filtering	0.92	0.91		0.92
[73]		Situational Information Filtering		0.99	0.66	0.74
[43]		Spatial Information Filtering	0.85	0.82		0.84

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100\% \tag{6}$$

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{7}$$

$$\text{Specificity/True Negative Rate (TNR)} = \frac{T_n}{T_n + F_p} \times 100\% \tag{8}$$

$$\text{Negative Predictive Value (NPV)} = \frac{T_n}{T_n + F_n}. \times 100\% \tag{9}$$

We also observed that Area Under the Receiver Operating Characteristic (ROC) curve value has been used by 6 articles. The ROC curve plots the values between sensitivity and (1-specificity). Sixty-four of the analysed articles presented their performance using more than one metric, while all of the remaining 19 used one metric only. Other metrics used by our analysed articles include Average Precision (AP), and Intersection over Union (IoU). Our analysis suggests that researchers primarily selected performance metrics based on the baseline work that they selected as a comparison for their results. Therefore, it is essential to use standard metrics so other researchers can compare and contrast results in future studies. Table 5 shows the best accuracy scores obtained for level-1 and and level-2 DR tasks in our taxonomy, revealing that across most tasks DL performs very well, with slightly lower success rates for sub-tasks such as *Damage evaluation* and *Spatial information filtering*.

$RQ_{4.2}$ What "baseline" Models have been Compared?

This question explores the benchmarks that have been chosen by the analysed articles. We observed that the vast majority of the analysed articles self-generated their own benchmark. Specifically, 35 of the studies evaluated the performance of their proposed approach against self-generated tests, while 25 evaluated DL approaches against classical ML approaches. We consider it likely that this is because, until recently, there have not been many DL-based approaches with which to compare. Moreover, the majority of the studies have not published their adopted models or code for future researchers to easily implement and

evaluate. Only 12 of the articles selected DL-methods proposed by previous research as baselines. We see that some benchmarks have also been compared in multiple articles as described in our online appendix [12].

RQ_5: What are the Underlying Challenges and Replicability of DL for DR Studies?

In RQ_5, we analyse the challenges researchers face in employing DL algorithms for DR studies and how well the current work can be adopted in future research. We aim to identify common challenges and provide future researchers with knowledge to better design future DL-based projects. Furthermore, we provide the details of research available for replication and reproduction in future research.

We observed that the challenges mostly depend on the data types and sources, including the following, which were extracted from 61 research articles:

1. *Data annotation*: Early studies using supervised approaches found very few publicly available annotated datasets. Therefore, they downloaded their own datasets and recruited people to annotate them. This took a massive amount of time and resources and delayed experiments. Furthermore, multi-label problems (one data item can belong to one or more informative categories), task subjectivity (difficulty in agreeing on one informative class), and conflicting annotation by human annotators were major issues. Even though many annotated datasets are available recently, data incompleteness and bias are common problems in processing DR data.
2. *High-level of noise:* Due to the high volume of heterogeneous data collected from social media platforms in the wake of disasters, the level of noise in the resulting data sets is extremely high (for example, spam, bots, data duplication). Furthermore, the content is informal, mostly using colloquial language, and very brief with casual acronyms and sometimes with non-literal language devices, like sarcasm, metaphors, and double entendre. Thus, it is challenging to train a DL model that can correctly interpret the intention of human expressions of this kind.
3. *High variability:* High variability in image quality resulting from different sensors and environmental conditions (for example, mist, cloud cover, and poor illumination) is challenging when applying DL models. Moreover, debris and damaged buildings look completely different depending on the disaster and structure of the building (e.g. concrete buildings, masonry buildings, or buildings made from natural materials), and are characterised by different features and patterns when captured in an image. As a result, the replicability of an already implemented solution for such a task is very low.
4. *Semantic segmentation*: Semantic segmentation of images to differentiate ground objects, such as roads and trees, from intact and damaged buildings, is a major challenge while using satellite, airborne and UAV imagery.

Despite these challenges, we observed that a very limited number of studies had made available their datasets, annotations, and implementation code for future research. For example, only 5 of the analysed articles made their resources publicly available. This trend results in researchers generating their own baseline and hence reducing research quality and the evolution of the field. Therefore, there is a considerable gap for researchers in adopting previous research as baselines.

Opportunities, Directions and Future Research Challenges

With the rapid change of climate and human-induced global warming, the variety and frequency of disasters have increased at a rate that has not happened before [28]. As a result, managing disasters while reducing their impacts on the communities and environment would be one of the main problems of the next decade. The increasing number of smart mobile devices and their embedded sensors enable the generation of a massive amount of heterogeneous data within a significantly shorter time than seen previously during disasters [1, 45]. Therefore, there is an immediate need for robust methods to automatically analyze and fuse such multimodal datasets and provide consolidated information to assist disaster management.

Data from different sources and formats bring complementary information regarding an event and lead to more robust inferences. Thus, future DL models will require analysis of heterogeneous, incomplete, and high-dimensional data sets to fill the missing information gaps in each data source or modality [98]. Multiple studies have explored the use of multimodal data for understanding the big picture of a disaster event [1, 3, 92, 99, 123]. However, more and more advanced DL approaches are required to solve core challenges in multimodal deep learning, such as missing data, dealing with different noise levels and effective fusing of heterogeneous data [17].

To address this problem, we identify that training data acquisition and preprocessing plays a major role when employing DL approaches. For example, large-scale human-annotated datasets are required to train DL algorithms to successfully predict the class label for unseen data. While a few annotated data repositories have been created (e.g.,

Table 6 Some association rules extracted from the analysed papers

Item	Support	Item	Confidence	Item	Lift
Supervised	0.94	Damage Assessment→ Remote Sensing	1.0	Multimodal, Twitter → CrisisMMD	4.50
CNN	0.70	Remote Sensing→ Image	1.0	Multimodal → CrisisMMD	3.46
Twitter	0.48	Multimodal, CrisisMMD→ Twitter	1.0	Remote Sensing → Image	2.24
image	0.45	Remote Sensing→ CNN	1.0	Remote Sensing, CNN → Image	2.24

CrisisNLP, CrisisMMD, and CrisisLex), more datasets are required to reflect temporal variations. Furthermore, there are still no large-scale benchmark datasets incorporating a variety of disaster data types except for CrisisMMD [10]. Therefore, the current research is mostly limited to small-scale home-grown datasets covering specific disaster types.

This leads to the next challenge of data irregularities occurring in datasets and which reduce a classifier's ability to learn from the data. The most common data irregularities include class imbalance, missing features, absent features, class skew and small disjuncts [29]. Class imbalance occurs when all classes present in a dataset do not have equal training instances. For example, datasets for classifying disaster-related social media posts have resulted in most non-related posts. Data-level methods, such as under-sampling techniques (e.g., Random Under-Sampling (RUS) [50]) and over-sampling techniques (e.g., Generative Adversarial Minority Oversampling (GAMO) [84] and Major-to-minor Translation (M2m) [54]), have been explored to mitigate the effects of class imbalance. Although researchers assume fully observed instances, practical datasets, however, contain missing features. Data imputation methods, model-based methods and more recently, DL methods have been proposed to handle missing features. A complete guide to methods that enable tackling these data irregularities is provided by Das et al. [29]. Even though methods to handle irregularities have been largely explored, more research is required as the velocity and variability of data generation accelerate.

Another key area is the variety characteristics of disasters that limit the reusability and generalizability of already trained DL algorithms. This means the variations of input data representations extracted during different disasters. Recent DL studies have focused on domain adaptation during learning where the distribution of the training data differs from the distribution of the test data [47]. Future research focus requires developing domain adaptation techniques for the DR domain.

According to the current trends, people will increasingly use social media platforms for disaster data acquisition, and dissemination, challenging the traditional media sources [40, 111, 118]. Therefore, crowd-sourced data will be more prominent in providing first-hand experiences of disaster scenes. However, responding organizations have concerns regarding the trustworthiness of user-generated content, a problem which is largely unsolved [23]. For example, fake news, misinformation, rumours, digital manipulation of images (e.g., deepfake [126]) and re-posting contents from previous events are a few challenges that future researchers will face to improve the integrity of social media content.

Another challenge in the DR domain is that previous research has largely explored the most common tasks, such as *Disaster damage assessment*, *Disaster event detection* and *Location reference identification*. However, there are other important DR tasks, including evacuation management, health and safety assurance, and critical infrastructure service, as illustrated in the Guidance of Emergency Response and Recovery [32]. These tasks have not yet been analyzed using DL approaches. Some possible reasons could be insufficient training datasets, lack of computational resources to store, manage, and process data, and inadequate accuracy of existing DL architectures. These underrepresented topics need further attention by DL researchers to better support DR tasks. Moreover, the accuracy of the output produced by DL algorithms is determined by a number of factors, including the optimization algorithm and the loss function used. Thus, further research is important in this area to find the correct combination of data, DL architecture, optimization algorithm, and loss function.

Results of the Association Rule Mining

This section discusses the interesting relationships discovered through our association rule mining task. We introduced the association rule mining process in Section "Association Rule Mining". Our goal is to identify hidden relationships between the values extracted from the articles for the attributes in the extraction form. The most highly scoring rules are listed in Table 6. We discuss the patterns that resulted in having higher "Support", "Confidence" and "Lift" values. However, all the associations are illustrated in our online appendix [12]. Our analysis highlights that CNN, Supervised, Image and Twitter have higher support values (> 0.45). This result indicates that the majority of studies discussed Image as data type, CNN as DL architecture, Supervised as learning type and Twitter as their data source.

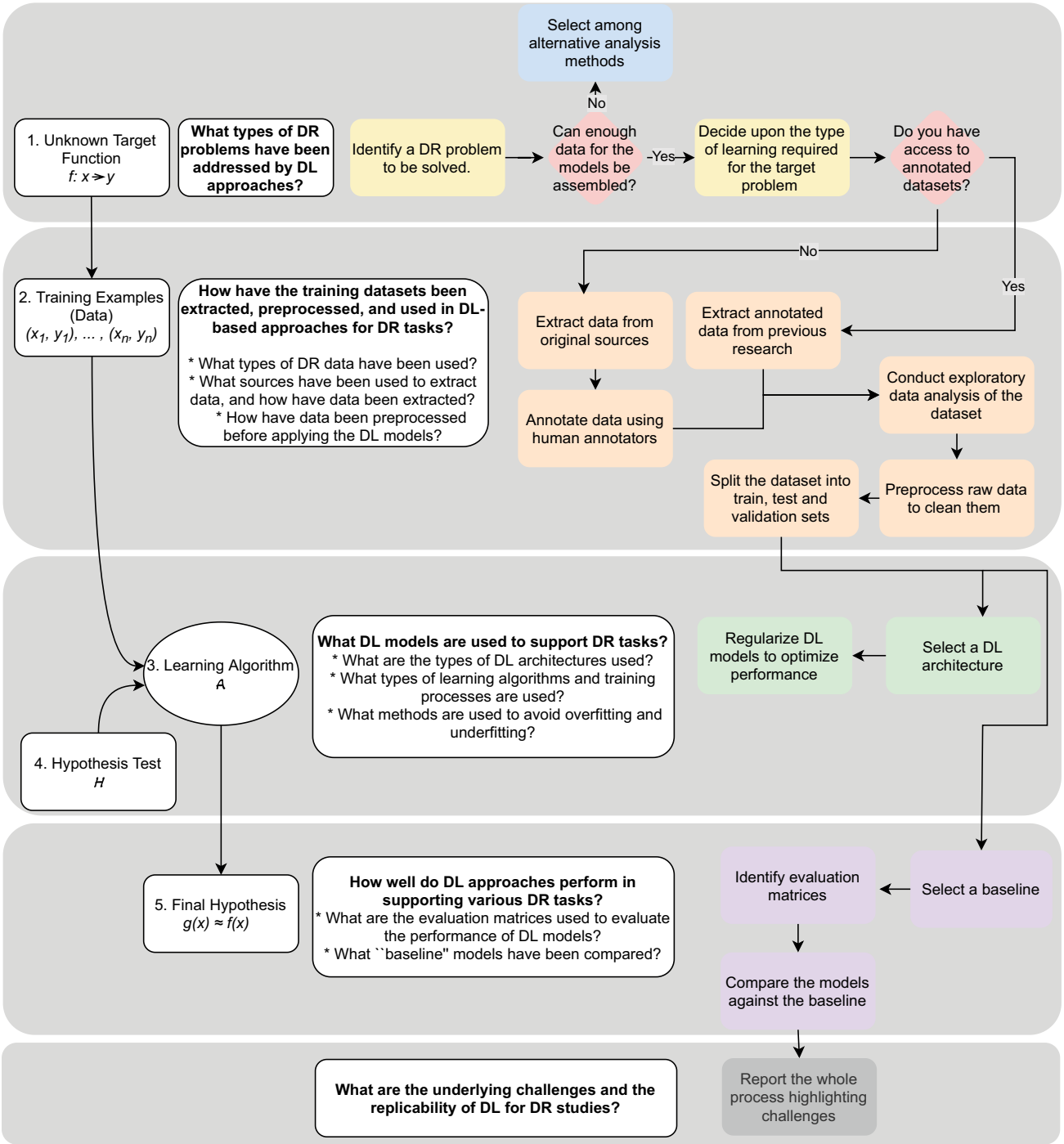


Fig. 13 Flowchart for conducting DL for DR research

Disaster Damage Assessment → Remote Sensing; *Remote Sensing → Image*; *Multimodal, CrisisMMD → Twitter* and *Remote Sensing, CNN→ Image* are some of the association rules having a confidence score of 1.0. This means that, for example, rule *Disaster Damage Assessment → Remote Sensing* implies that the pattern appears in 100% of the analysed articles. Similarly, all the research that used Remote Sensing as the data extraction method analysed Image as their data source.

The highest lift score of 4.5 resulted for the *multimodal, Twitter→ CrisisMMD* rule. This means that when researchers used multimodal as their data type and Twitter as the

Data Source, CrisisMMD has commonly been the data extraction method. Furthermore, *multimodal→CrisisMMD, Twitter*; *Remote Sensing → Disaster Damage Assessment, Image*; *Image→Remote Sensing, CNN* rules were among the other high lift values. Interestingly, we observed rules such as *Twitter→CNN*; *CNN → text* and *text, Twitter → CNN* having a "Lift" score of less than 1. This indicates a negative relationship between the parameter values. For example, it is very unlikely that research that used Text as the Data Type and CNN as the DL architecture. All these association rules provide future researchers a guide to select parameters in a DL-based project, such as data sources, learning algorithms, and learning type.

Flowchart and Guidelines for Applying DL in Future DR Research

In this section, we provide a flowchart and guidelines for conducting future work using DL for DR tasks based on the findings of our SLR. Figure 13 shows how we have mapped the components of learning into RQs and then as the steps in the flowchart. The extracted flowchart is a general one based on the 83 analyzed papers. However, more specific details can be added to it based on the DR task to be solved.

After identifying the DR problem to be addressed, researchers should consider whether DL is a suitable approach. That decision can be made partly based on whether it is possible to obtain or create the necessary data. If enough data can be obtained, the researcher can select either supervised, unsupervised and semi-supervised learning methods. We discussed these methods in the Section "$RQ_{3.2}$ What Training Processes are Used to Optimize DL Models?". If the identified problem can be better solved using a supervised approach, the next step is to decide where the annotated datasets can be obtained, or whether raw data must be annotated. Data annotation is generally labour intensive and time-consuming, and therefore, the researcher can hire paid workers or arrange volunteers based on budget and availability. We have discussed the annotated data sources and annotation methods in the Sections $RQ_{2.2}$ What Sources have been Used to Extract Data, and How Have Data Been Extracted? and $RQ_{2.3}$ How have data been Preprocessed Before Applying the DL Models?. Once the dataset is ready, the researcher should conduct an exploratory analysis to identify the nature of this raw data. This analysis provides the researcher with an overview including the size, distribution, and characteristics of the data. Proper understanding of raw data provides guidelines for the design of the preprocessing steps, which have to be well reported to enable replication. This includes outlining all the steps involved, including the normalization processes and data augmentation strategies.

After the data filtering and cleaning steps, the researcher should identify the learning algorithm, and DL architecture. The researcher should report the details of the DL architecture, including the type of layer (e.g., embedding, dropout and soft-max), number of layers, filters, and learning rate. Furthermore, all necessary details regarding optimizers, loss function and hyper-parameter tuning, have to be reported to enable replication. The information regarding training, such as number of iterations (epochs), strategies combating overfitting and underfitting, training time, computing environment, special computing resources (e.g., GPUs, high-performance computing) and platforms used (e.g., Google Colaborotory) should also be explained (see Section "RQ_3: What DL Models are Used to Support DR Tasks?").

Finally, the researcher should report the results compared to the selected "baseline model". If the researchers used their own dataset, they must first implement the baseline against their data to compare the results. Any limitations and challenges encountered while applying DL models should also be discussed to provide guidance for future researchers in designing DL-based approaches for DR tasks. Furthermore, researchers can support the quality and the future of the DR research field by making publicly available the datasets, annotations, and DL architectures.

Conclusion

This study has presented a systematic literature review of DL in DR research. We started by identifying RQs for the analysis according to the components of learning described by Abu Moftha [121]. Then, a data extraction form with 15 attributes was created to extract answers for the questions from the selected articles. Finally, we used the KDD process to identify relationships among different attributes of the extracted data. The answers to the research questions indicate that, while some DR tasks have received much investigation, others have received less attention. Furthermore, there are multiple challenges while collecting, annotating, and preprocessing datasets for DL tasks. However, researchers have achieved better performance than traditional methods when using DL methods for DR tasks despite these challenges.

This research has identified opportunities, future research challenges, and many directions for further investigation. For example, multiple DR tasks are yet to be studied using DL approaches, such as *evacuation management* and *critical infrastructure services*. Moreover, we highlighted the need for new annotated multimodal datasets targeted at DR concerns. Some of the future research challenges are handling data irregularities, improving the integrity of social media data, and developing generalizable DL approaches across multiple disasters. Additionally, data preprocessing, DL architecture selection, word embeddings and hyperparameter tuning are areas of further exploration. Finally, we emphasized the importance of comprehensive reporting and making implemented DL methodologies publicly available for the advancement of the DL in the DR area.

Appendix A: Glossary of Terms

Table 7 shows the expansions of abbreviated terms used in the paper.

Table 7 Glossary of Terms

Term	Definition
Adam	Adaptive Moment Estimation
AE	AutoEncoder
AI	Artificial Intelligence
AP	Avergae Precision
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bi-directional LSTM
CCTV	Closed-Circuit Television
CNN	Convolutional Neural Network
DANN	Domain Adversarial Neural Network
DL	Deep Learning
DR	Disaster Response
GPU	Graphical Processing Units
IoU	Intersection over Union
KDD	Knowledge Discovery in Databases
LiDAR	Light Detection and Ranging
LSTM	Long Short-Term Memory Network
ML	Machine Learning
NLP	Natural Language Processing
OOV	Out of Vocabulary
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized BERT Pre-training Approach
ROC	Receiver Operating Characteristic
RQ	Research Question
SLR	Systematic Literature Review
UAV	Unmanned Aerial Vehicle
URL	Uniform Resource Locator

Appendix B: Publication Venues

Table 8 provides article publication venues that are also listed in our online appendix [12].

Table 8 Article Publication Venues

Journal/ Conference Name	Abbreviation
AAAI Conference on Artificial Intelligence	AAAI
Advanced Engineering Informatics	AEI
Applied Imagery Pattern Recognition Workshop	AIPr
Advances in Intelligent Systems and Computing	AISC
Annals of Operations Research	AOR
arXiv	arXiv
Computer-Aided Civil and Infrastructure Engineering	CACIE
International Conference on Communication Systems and Networks	COMSNETS
Conference on Computer Vision and Pattern Recognition	CVPR
International Electronics Symposium on Knowledge Creation and Intelligent Computing	DSAA
Decision Support Systems	DSS
Intelligent Computing in Engineering	EG-ICE
Information and Communication Technologies for Disaster Management	ICT-DM
Institute of Electrical and Electronics Engineers	IEEE
International Electronics Symposium on Knowledge Creation and Intelligent Computing	IES-KCIC
International Geoscience and Remote Sensing Symposium	IGARSS
International Journal of Digital Earth	IJDE
International Journal of Disaster Risk Reduction	IJDRR
International Journal of Distributed Sensor Networks	IJDSN
International Journal of Innovative Technology and Exploring Engineering	IJITEE
IOP Conference Series: Materials Science and Engineering	IOP
Information Systems for Crisis Response And Management	ISCRAM
Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences	ISPRS
Innovations in Systems and Software Engineering	ISSE
International Semantic Web Conference	ISW
Journal of Ambient Intelligence and Humanized Computing	JAIHC
Journal of Applied Remote Sensing	JARS
Journal on Computing and Cultural Heritage	JCCH
International Conference on Mobile Data Management	MDM
MediaEval	MediaEval
Conference on Multimedia Information Processing and Retrieval	MIPR
Multimedia Tools and Applications	MTA
Neurocomputing	Neurocomputing
Procedia Computer Science	PCS
Progress in Disaster Science	PDS
Procedia Engineering	Procedia Engineering
Remote Sensing	Remote Sensing
Sadhana - Academy Proceedings in Engineering Sciences	SADHANA
Structural Control and Health Monitoring	SCHM
Sensors	Sensors
ACM Symposium on Applied Computing	SIGAPP
International Conferences on Advances in Geographic Information Systems	SIGSPATIAL
Signal Processing: Image Communication	SPIC
World Wide Web	WWW

Declarations

Conflicts interest The authors declare no conflicts interest.

References

1. Abavisani M, Wu L, Hu S, Tetreault J, Jaimes A. Multimodal categorization of crisis events in social media. arXiv (2020).
2. Acerbo FS, Rossi C. Filtering informative tweets during emergencies: a machine learning approach. In: I-TENDER 2017—Proceedings of the 2017 1st CoNEXT Workshop on ICT Tools for Emergency Networks and DisastEr Relief pp. 1–6 (2017). https://doi.org/10.1145/3152896.3152897.
3. Agarwal M, Leekha M, Sawhney R, Shah RR. Crisis-DIAS: towards multimodal damage analysis—deployment, challenges and assessment. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020;34(01):346–53. https://doi.org/10.1609/aaai.v34i01.5369.
4. Ahmad K, Pogorelov K, Riegler M, Ostroukhova O, Halvorsen P, Conci N, Dahyot R. Automatic detection of passable roads after floods in remote sensed and social media data. Signal Process Image Commun. 2019;74((December 2018)):110–8. https://doi.org/10.1016/j.image.2019.02.002.
5. Aipe A, Ekbal A, S, MN, Kurohashi S. Linguistic feature assisted deep learning approach towards multi-label classification of crisis related tweets. In: Boersma K, Tomaszewski BM (eds.) Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018. ISCRAM Association (2018). http://idl.iscram.org/files/alanaipe/2018/1592_AlanAipe_etal2018.pdf.
6. Akter S, Wamba SF. Big data and disaster management: a systematic review and agenda for future research. Ann Oper Res. 2017. https://doi.org/10.1007/s10479-017-2584-2.
7. Alam F, Hassan Z, Ahmad K, Gul A, Reiglar M, Conci N, Al-Fuqaha A. Flood detection via twitter streams using textual and visual features. arXiv 2020; p. 4–6
8. Alam F, Imran M, Ofli F. Image4Act: online social media image processing for disaster response. In: Proceedings of the 2017 IEEE/ACM International Conference on advances in social networks analysis and mining, ASONAM 2017 2017;601–604. https://doi.org/10.1145/3110025.3110164.
9. Alam F, Joty S, Imran M. Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. In: Twelfth International AAAI Conference on Web and Social Media (2018). Accessed 10 May 2021.
10. Alam F, Ofli F, Imran M. Crisismmd: multimodal twitter datasets from natural disasters. In: Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25–28, 2018, pp. 465–473. AAAI Press 2018. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17816.
11. Alexander DE. Social media in disaster risk reduction and crisis management. Sci Eng Ethics. 2014;20(3):717–33.
12. Algiriyage N, Doyle EEH, Stock K, Johnston D. Data analysis details of the systematic literature review of dl for dr. https://github.com/mu-clab/DLforDR.
13. Amit SNKB, Aoki Y. Disaster detection from aerial imagery with convolutional neural network. In: Proceedings—International Electronics Symposium on knowledge creation and intelligent computing, IES-KCIC 2017, 2017—January (July 2018), p. 239–245 (2017). https://doi.org/10.1109/KCIC.2017.8228593.
14. Anbarasan M, Muthu B, Sivaparthipan C, Sundarasekar R, Kadry S, Krishnamoorthy S, Dasel AA, et al. Detection of flood disaster system based on iot, big data and convolutional deep neural network. Comput Commun. 2020;150:150–7. https://doi.org/10.1016/j.comcom.2019.11.022.
15. Arif Amin MA, Ali AA, Rahman AK. Visual attention-based comparative study on disaster detection from social media images. Innov Syst Softw Eng. 2020;16(3–4):309–19. https://doi.org/10.1007/s11334-020-00368-1. Accessed 10 May 2021.
16. Attari N, Ofli F, Awad M, Lucas J, Chawla S. Nazr-CNN: Fine-grained classification of UAV imagery for damage assessment. In: Proceedings—2017 International Conference on data science and advanced analytics, DSAA 2017, 2018-January, p. 50–59 (2017). https://doi.org/10.1109/DSAA.2017.72.
17. Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41(2):423–43. Accessed 10 May 2021.
18. Bansal S. A Mutli-task mutlimodal framework for tweet classification based on CNN (Grand Challenge). In: Proceedings—2020 IEEE 6th International Conference on multimedia big data, BigMM 2020; pp. 456–460 (2020). https://doi.org/10.1109/BigMM50055.2020.00075.
19. Basu M, Shandilya A, Khosla P, Ghosh K, Ghosh S. Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations. IEEE Trans Comput Soc Syst. 2019;6(3):604–18. https://doi.org/10.1109/TCSS.2019.2914179.
20. Bejiga MB, Zeggada A, Nouffidj A, Melgani F. A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery. Remote Sens. 2017. https://doi.org/10.3390/rs9020100.
21. Burel G, Alani H. Crisis event extraction service (CREES)—automatic detection and classification of crisis-related content on social media. In: Proceedings of the International ISCRAM Conference 2018; 2018-May, pp. 597–608.
22. Burel G, Saif H, Alani H. Semantic wide and deep learning for detecting crisis-information categories on social media. In: International Semantic Web Conference, 2017;138–155. Springer. https://doi.org/10.1007/978-3-319-68288-4_9.
23. Castillo C. Big crisis data: social media in disasters and time-critical situations. Cambridge: Cambridge University Press; 2016.
24. Cha YJ, Choi W, Büyüköztürk O. Deep learning-based crack damage detection using convolutional neural networks. Comput-Aided Civ Infrastruct Eng. 2017;32(5):361–78. https://doi.org/10.1111/mice.12263.
25. Chaudhuri N, Bose I. Exploring the role of deep neural networks for post-disaster decision support. Decis Support Syst. 2020;130((July 2019)):130. https://doi.org/10.1016/j.dss.2019.113234.
26. Chen F, Yu B. Earthquake-induced building damage mapping based on multi-task deep learning framework. IEEE Access. 2019;7:181396–404. https://doi.org/10.1109/ACCESS.2019.2958983.
27. Cheng CS, Behzadan AH, Noshadravan A. Deep learning for post-hurricane aerial damage assessment of buildings. Comput-Aided Civ Infrastruct Eng. 2021;1:16. https://doi.org/10.1111/mice.12658.
28. Climate change: how do we know? https://climate.nasa.gov/evidence/. Accessed 29 Oct 2021.
29. Das S, Datta S, Chaudhuri BB. Handling data irregularities in classification: Foundations, trends, and future challenges. Pattern Recognit. 2018;81:674–93. https://doi.org/10.1016/j.patcog.2018.03.008.
30. Devaraj A, Murthy D, Dontula A. Machine-learning methods for identifying social media-based requests for urgent help during hurricanes. Int J Disaster Risk Reduct. 2020. https://doi.org/10.1016/j.ijdrr.2020.101757.
31. Duarte D, Nex F, Kerle N, Vosselman G. Satellite image classification of building damages using airborne and satellite image

samples in a deep learning approach. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci. 2018. https://doi.org/10.5194/isprs-annals-IV-2-89-2018.
32. Emergency response and recovery. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/253488/Emergency_Response_and_Recovery_5th_edition_October_2013.pdf (2013). Accessed 30 Apr 2021.
33. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37–54. https://doi.org/10.1609/aimag.v17i3.1230.
34. Gao W, Li L, Zhu X, Wang Y. Detecting disaster-related tweets via multimodal adversarial neural network. IEEE Multimed. 2020;27(4):28–37. https://doi.org/10.1109/MMUL.2020.3012675.
35. Ghaffarian S, Kerle N, Pasolli E, Arsanjani JJ. Post-disaster building database updating using automated deep learning: an integration of pre-disaster OpenStreetMap and multi-temporal satellite data. Remote Sens. 2019;11(20):1–20. https://doi.org/10.3390/rs11202427.
36. Ghosh Mondal T, Jahanshahi MR, Wu RT, Wu ZY. Deep learning-based multi-class damage detection for autonomous post-disaster reconnaissance. Struct Control Health Monit. 2020;27(4):1–15. https://doi.org/10.1002/stc.2507.
37. Gomez C, Purdie H. Uav-based photogrammetry and geocomputing for hazards and disaster risk monitoring-a review. Geoenviron Disasters. 2016;3(1):23. https://doi.org/10.1186/s40677-016-0060-y.
38. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Inf Libr J. 2009;26(2):91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x.
39. Hao H, Wang Y. Leveraging multimodal social media data for rapid disaster damage assessment. Int J Disaster Risk Reduct. 2020. https://doi.org/10.1016/j.ijdrr.2020.101760.
40. Hao H, Wang Y. Leveraging multimodal social media data for rapid disaster damage assessment. Int J Disaster Risk Reduct. 2020. https://doi.org/10.1016/j.ijdrr.2020.101760.
41. Hartawan DR, Purboyo TW, Setianingsih C. Disaster victims detection system using convolutional neural network (CNN) method. In: Proceedings—2019 IEEE International Conference on Industry 4.0, artificial intelligence, and communications technology, IAICT 2019 2019;105–111. https://doi.org/10.1109/ICIAICT.2019.8784782.
42. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci. 2004;44(1):1–12.
43. Hernandez-Suarez A, Sanchez-Perez G, Toscano-Medina K, Perez-Meana H, Portillo-Portillo J, Sanchez V, Villalba LJG. Using twitter data to monitor natural disaster social dynamics: a recurrent neural network approach with word embeddings and kernel density estimation. Sensors (Switzerland). 2019. https://doi.org/10.3390/s19071746.
44. Hezaveh MM, Kanan C, Salvaggio C. Roof damage assessment using deep learning. In: Proceedings—Applied Imagery Pattern Recognition Workshop 2018; 2017-October, p. 6403–6408. https://doi.org/10.1109/AIPR.2017.8457946.
45. Huang X, Li Z, Wang C, Ning H. Identifying disaster related social media for rapid response: a visual-textual fused CNN architecture. Int J Digit Earth. 2020;13(9):1017–39. https://doi.org/10.1080/17538947.2019.1633425.
46. Huang X, Wang C, Li Z, Ning H. A visual-textual fused approach to automated tagging of flood-related tweets during a flood event. Int J Digit Earth. 2019;12(11):1248–64. https://doi.org/10.1080/17538947.2018.1523956.
47. Imran M, Mitra P, Srivastava J. Cross-language domain adaptation for classifying crisis-related short messages. In: Tapia AH, Antunes P, Bañuls VA, Moore KA, de Albuquerque JP (eds.) 13th Proceedings of the International Conference on information systems for crisis response and management, Rio de Janeiro, Brasil, May 22–25, 2016. ISCRAM Association 2016. http://idl.iscram.org/files/muhammadimran/2016/1396_MuhammadImran_etal2016.pdf.
48. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on machine learning, 2015; p. 448–456. PMLR.
49. Jabbar H, Khan RZ. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Commun Instrum Devices Comput Sci. 2015; p. 163–72.
50. Japkowicz N. The class imbalance problem: significance and strategies. In: Proceedings of the International Conference on Artificial Intelligence, vol. 56. Citeseer 2000. Accessed 10 May 2021.
51. Jones S, Saniie J. Using deep learning and satellite imagery to assess the damage to civil structures after natural disasters. In: IEEE International Conference on electro information technology. 2019;189–93. https://doi.org/10.1109/EIT.2019.8833724.
52. Kabir MY, Madria S. A deep learning approach for tweet classification and rescue scheduling for effective disaster management. Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL. 2019;269–78. https://doi.org/10.1145/3347146.3359097.
53. Khan SH, He X, Porikli F, Bennamoun M. Forest change detection in incomplete satellite images with deep neural networks. IEEE Trans Geosci Remote Sens. 2017;55(9):5407–23. https://doi.org/10.1109/TGRS.2017.2707528.
54. Kim J, Jeong J, Shin J. M2m: Imbalanced classification via major-to-minor translation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020, pp. 13893–13902. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01391, https://openaccess.thecvf.com/content_CVPR_2020/html/Kim_M2m_Imbalanced_Classification_via_Major-to-Minor_Translation_CVPR_2020_paper.html.
55. Kruspe A, Kersten J, Klan F. Detection of informative tweets in crisis events. In: Natural Hazards and Earth System Sciences (NHESS) 2021.
56. Kumar A, Singh JP. Location reference identification from tweets during emergencies: a deep learning approach. Int J Disaster Risk Reduct. 2019;33:365–75. https://doi.org/10.1016/j.ijdrr.2018.10.021.
57. Kumar A, Singh JP. Disaster severity prediction from Twitter images. Adv Intell Syst Comput. 2021;1279(December 2020):65–73. https://doi.org/10.1007/978-981-15-9290-4_7.
58. Kumar A, Singh JP, Dwivedi YK, Rana NP. A deep multi-modal neural network for informative Twitter content classification during emergencies. 0123456789. Ann Oper Res. 2020. https://doi.org/10.1007/s10479-020-03514-x.
59. Kumar P, Ofli F, Imran M, Castillo C. Detection of disaster-affected cultural heritage sites from social media images using deep learning techniques. J Comput Cult Herit. 2020;13:3. https://doi.org/10.1145/3383314.
60. Kundu S, Srijith PK, Desarkar MS. Classification of short-texts generated during disasters: a deep neural network based approach. In: Proceedings of the 2018 IEEE/ACM International Conference on advances in social networks analysis and mining, ASONAM 2018 2018; pp. 790–793. https://doi.org/10.1109/ASONAM.2018.8508695.
61. Laghrabli S, Benabbou L, Berrado A. A new methodology for literature review analysis using association rules mining. In: 10th International Conference on intelligent systems: theories and applications, SITA 2015, Rabat, Morocco, October 20–21,

2015, 2015; pp. 1–6. IEEE . https://doi.org/10.1109/SITA.2015.7358394.
62. Li X, Caragea C, Caragea D, Imran M, Ofli F. Identifying disaster damage images using a domain adaptation approach. Proceedings of the International ISCRAM Conference 2019; 2019-May (May 2019), pp. 633–645.
63. Li X, Caragea D. Improving Disaster-related Tweet Classification with a Multimodal Approach. Social Media for Disaster Response and Resilience Proceedings ofthe 17th ISCRAM Conference (May), 2020;893–902
64. Li X, Caragea D, Zhang H, Imran M. Localizing and quantifying damage in social media images. In: Proceedings of the 2018 IEEE/ACM International Conference on advances in social networks analysis and mining, ASONAM 2018 2018; pp. 194–201. https://doi.org/10.1109/ASONAM.2018.8508298.
65. Li Y, Caragea C, Park S, Caragea D, Tapia A. Sympathy detection in disaster Twitter data. In: Proceedings of the International ISCRAM Conference. 2019;788–798. http://idl.iscram.org/files/yingjieli/2019/1899_YingjieLi_etal2019.pdf.
66. Li Y, Ye S, Bartoli I. Semisupervised classification of hurricane damage from postevent aerial imagery using deep learning. J Appl Remote Sens. 2018;12(4):1–13. https://doi.org/10.1117/1.JRS.12.045008.
67. Li Y, Ye S, Bartoli I. Semisupervised classification of hurricane damage from postevent aerial imagery using deep learning. J Appl Remote Sens. 2018;12(04):1. https://doi.org/10.1117/1.jrs.12.045008.
68. Liang X. Image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with Bayesian optimization. Comput-Aided Civ Infrastruct Eng. 2019;34(5):415–30. https://doi.org/10.1111/mice.12425.
69. Liu Y, Wu L. Geological disaster recognition on optical remote sensing images using deep learning. Proc Comput Sci. 2016;91(Itqm):566–75. https://doi.org/10.1016/j.procs.2016.07.144.
70. Liu Y, Wu L. High performance geological disaster recognition using deep learning. Proc Comput Sci. 2018;139:529–36. https://doi.org/10.1016/j.procs.2018.10.237.
71. Lohumi K, Roy S. Automatic detection of flood severity level from flood videos using deep learning models. In: 2018 5th International Conference on information and communication technologies for disaster management, ICT-DM 2018 2019; pp. 1–7. https://doi.org/10.1109/ICT-DM.2018.8636373.
72. Luna S, Pennock MJ. Social media applications and emergency management: a literature review and research agenda. Int J Disaster Risk Reduct. 2018;28:565–77. https://doi.org/10.1016/j.ijdrr.2018.01.006.
73. Madichetty S, Muthukumarasamy S. Detection of situational information from Twitter during disaster using deep learning models. Sadhana Acad Proc Eng Sci. 2020;45(1):1–13. https://doi.org/10.1007/s12046-020-01504-0.
74. Madichetty S, Muthukumarasamy S, Jayadev P. Multi-modal classification of Twitter data during disasters for humanitarian response. J Ambient Intell Humaniz Comput. 2021. https://doi.org/10.1007/s12652-020-02791-5.
75. Madichetty S, Sridevi M. Detecting informative tweets during disaster using deep neural networks. In: 2019 11th International Conference on communication systems and networks, COMSNETS 2019. 2019;2061:709–13. https://doi.org/10.1109/COMSNETS.2019.8711095.
76. Madichetty S, Sridevi M. Classifying informative and non-informative tweets from the twitter by adapting image features during disaster. Multimed Tools Appl. 2020;79(39–40):28901–23. https://doi.org/10.1007/s11042-020-09343-1.
77. Madichetty S, Sridevi M. A stacked convolutional neural network for detecting the resource tweets during a disaster. Multimed Tools Appl. 2021;80(3):3927–49. https://doi.org/10.1007/s11042-020-09873-8.
78. Mangalathu S, Burton HV. Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions. Int J Disaster Risk Reduct. 2019. https://doi.org/10.1016/j.ijdrr.2019.101111.
79. Miura H, Aridome T, Matsuoka M. Deep learning-based identification of collapsed, non-collapsed and blue tarp-covered buildings from post-disaster aerial images. Remote Sens. 2020. https://doi.org/10.3390/rs12121924.
80. Moechammad S, Cahya R, Berkah ANA. Detecting body parts from natural disaster victims using You Only Look Once (YOLO). IOP Conf Ser Mater Sci Eng. 2021. https://doi.org/10.1088/1757-899x/1073/1/012062.
81. Mohanty SD, Biggers B, Sayedahmed S, Pourebrahim N, Goldstein EB, Bunch R, Chi G, Sadri F, McCoy TP, Cosby A. A multi-modal approach towards mining social media data during natural disasters - A case study of Hurricane Irma. Int J Disaster Risk Reduct. 2021;54(July 2020):102032. https://doi.org/10.1016/j.ijdrr.2020.102032.
82. Mouzannar H, Rizk Y, Awad M. Damage identification in social media posts using multimodal deep learning. In: Proceedings of the International ISCRAM Conference 2018; 2018-May(May), pp. 529–543.
83. Muhammad K, Ahmad J, Baik SW. Early fire detection using convolutional neural networks during surveillance for effective disaster management. Neurocomputing. 2018;288:30–42. https://doi.org/10.1016/j.neucom.2017.04.083.
84. Mullick SS, Datta S, Das S. Generative adversarial minority oversampling. In: 2019 IEEE/CVF International Conference on computer vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019, pp. 1695–1704. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00178.
85. Naga Pavan Srivathsav C, Anitha K, Anvitha K, Maneesha B, Sagar Imambi S. Detection of disaster affected regions based on change detection using deep architecture. Int J Innov Technol Explor Eng. 2019;8(5):124–8.
86. Neppalli VK, Caragea C, Caragea D. Deep neural networks versus Naive Bayes classifiers for identifying informative tweets during disasters. In: Boersma K, Tomaszewski BM (eds.) Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20–23, 2018. ISCRAM Association 2018. http://idl.iscram.org/files/venkatakishoreneppalli/2018/1589_VenkataKishoreNeppalli_etal2018.pdf.
87. Neppalli VK, Caragea C, Caragea D. Deep neural networks versus Naïve Bayes classifiers for identifying informative tweets during disasters. In: Proceedings of the International ISCRAM Conference 2018; 2018-May (May), pp. 677–686.
88. Nex F, Duarte D, Tonolo FG, Kerle N. Structural building damage detection with deep learning: assessment of a state-of-the-art CNN in operational conditions. Remote Sens. 2019. https://doi.org/10.3390/rs11232765.
89. Nguyen DT, Ofli F, Imran M, Mitra P. Damage assessment from social media imagery data during disasters. In: Proceedings of the 2017 IEEE/ACM International Conference on advances in social networks analysis and mining, ASONAM 2017 2017;569–76. https://doi.org/10.1145/3110025.3110109. Accessed 10 May 2021.
90. Nguyen VQ, Anh TN, Yang HJ. Real-time event detection using recurrent neural network in social sensors. Int J Distrib Sens Netw. 2019;(15)6. https://doi.org/10.1177/1550147719856492.
91. Nunavath V, Goodwin M. The role of artificial intelligence in social media big data analytics for disaster management-initial results of a systematic literature review. In: 2018 5th International Conference on information and communication technologies for

disaster management (ICT-DM), 2018;1–4. IEEE. https://doi.org/10.1109/ICT-DM.2018.8636388.
92. Ofli F, Alam F, Imran M. Analysis of social media data using multimodal deep learning for disaster response 2020;**1**(May 2020). arXiv: 2004.11838.
93. Padhee S, Saha TK, Tetreault J, Jaimes A. Clustering of social media messages for humanitarian aid response during crisis. arXiv 2020.
94. Parilla-Ferrer BE, Fernandez PL, Ballena JT. Automatic classification of disaster-related tweets. In: Proc. International Conference on innovative engineering technologies (ICIET), 2014; vol. 62.
95. Pi Y, Nath ND, Behzadan AH. Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. Adv Eng Inf. 2020. https://doi.org/10.1016/j.aei.2019.101009.
96. Pi Y, Nath ND, Behzadan AH. Disaster impact information retrieval using deep learning object detection in crowdsourced drone footage. In: EG-ICE 2020 Workshop on Intelligent Computing in Engineering, Proceedings 2020; pp. 134–143.
97. Pogrebnyakov N, Maldonado EA. Identifying emergency stages in facebook posts of police departments with convolutional and recurrent neural networks and support vector machines. In: 2017 IEEE International Conference on Big Data (IEEE BigData),2017;4343–52. IEEE Computer Society (2017). https://doi.org/10.1109/BigData.2017.8258464.
98. Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MEP, Shyu M, Chen S, Iyengar SS. A survey on deep learning: algorithms, techniques, and applications. ACM Comput Surv. 2019;51(5):92:1-92:36. https://doi.org/10.1145/3234150.
99. Pouyanfar S, Tao Y, Tian H, Chen SC, Shyu ML. Multimodal deep learning based on multiple correspondence analysis for disaster management. World Wide Web. 2018. https://doi.org/10.1007/s11280-018-0636-4.
100. Pranesh RR, Shekhar A, Kumar A. Exploring multimodal features and fusion strategies for analyzing disaster tweets.
101. Prasanna R, Huggins TJ. Factors affecting the acceptance of information systems supporting emergency operations centres. Comput Hum Behav. 2016;57:168–81. https://doi.org/10.1016/j.chb.2015.12.013.
102. Prasanna R, Yang L, King M. Guidance for developing human-computer interfaces for supporting fire emergency response. Risk Manag. 2013;15(3):155–79. https://doi.org/10.1057/rm.2013.3.
103. Priya S, Bhanu M, Dandapat SK, Ghosh K, Chandra J. TAQE: tweet retrieval-based infrastructure damage assessment during disasters. IEEE Trans Comput Soc Syst. 2020;7(2):389–403. https://doi.org/10.1109/TCSS.2019.2957208.
104. Public feed api. https://developers.facebook.com/docs/public_feed/. Accessed 10 May 2021.
105. Qadir J, Ali A, ur Rasool R, Zwitter A, Sathiaseelan A, Crowcroft J. Crisis analytics: big data-driven crisis response. J Int Hum Action. 2016;1(1):12. https://doi.org/10.1186/s41018-016-0013-9.
106. Rahnemoonfar M, Murphy R, Miquel MV, Dobbs D, Adams A. Flooded area detection from UAV images based on densely connected recurrent neural networks. In: International Geoscience and Remote Sensing Symposium (IGARSS) 2018; 2018-July, 1788–1791. https://doi.org/10.1109/IGARSS.2018.8517946.
107. Rice L, Wong E, Kolter JZ. Overfitting in adversarially robust deep learning. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, *Proceedings of Machine Learning Research*, vol. 119, pp. 8093–8104. PMLR 2020. http://proceedings.mlr.press/v119/rice20a.html.
108. Rizk Y, Awad M, Jomaa HS, Castillo C. A computationally efficient multi-modal classification approach of disaster-related Twitter images. In: Proceedings of the ACM Symposium on applied computing Part. 2019;F1477(January):2050–9. https://doi.org/10.1145/3297280.3297481.
109. Robertson BW, Johnson M, Murthy D, Smith WR, Stephens KK. Using a combination of human insights and 'deep learning' for real-time disaster communication. Progress Disaster Sci. 2019;2:100030. https://doi.org/10.1016/j.pdisas.2019.100030.
110. Sadiq AM, Ahn H, Choi YB. Human sentiment and activity recognition in disaster situations using social media images based on deep learning. Sensors. 2020;20(24):7115. https://doi.org/10.3390/s20247115.
111. Sahoh B, Choksuriwong A. Smart emergency management based on social big data analytics: Research trends and future directions. In: Proceedings of the 2017 International Conference on Information Technology, 2017;1–6. https://doi.org/10.1145/3176653.3176657.
112. Seydi ST, Rastiveis H. A deep learning framework for roads network damage assessment using post-earthquake lidar data. Int Arch Photogramm Remote Sens Spatial Inf Sci ISPRS Arch. 2019;42(4/W18):955–61. https://doi.org/10.5194/isprs-archives-XLII-4-W18-955-2019.
113. Shams S, Goswami S, Lee K. Deep learning-based spatial analytics for disaster-related tweets: an experimental study. In: Proceedings—IEEE International Conference on mobile data management 2019-June (Mdm), 2019;337–342. https://doi.org/10.1109/MDM.2019.00-40.
114. Shin H, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura DJ, Summers RM. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. CoRR **abs/1602.03409** (2016). arxiv:1602.03409.
115. Sit MA, Koylu C, Demir I. 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. Int J Digit Earth. 2019;12(11):1205–29. https://doi.org/10.1080/17538947.2018.1563219.
116. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;**15**(1):1929–58. http://dl.acm.org/citation.cfm?id=2670313.
117. Sublime J, Kalinicheva E. Automatic post-disaster damage mapping using deep-learning techniques for change detection: case study of the Tohoku tsunami. Remote Sens. 2019. https://doi.org/10.3390/rs11091123.
118. Sun W, Bocchini P, Davison BD. Applications of artificial intelligence for disaster management. Dordrecht: Springer Netherlands; 2020. https://doi.org/10.1007/s11069-020-04124-3 (**0123456789**). Accessed 10 May 2021.
119. Tatulli E, Hueber T. Feature extraction using multimodal convolutional neural networks for visual speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2971–2975. IEEE (2017). https://doi.org/10.1109/ICASSP.2017.7952701.
120. Tian H, Zheng HC, Chen S. Sequential deep learning for disaster-related video classification. IEEE 1st Conference on multimedia

information processing and retrieval (MIPR). 2018;106–111. IEEE 2018.https://doi.org/10.1109/MIPR.2018.00026.
121. The learning problem. http://work.caltech.edu/slides/slides01.pdf. Accessed 10 May 2021.
122. Vetrivel A, Gerke M, Kerle N, Nex F, Vosselman G. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. ISPRS J Photogramm Remote Sens. 2018;140:45–59. https://doi.org/10.1016/j.isprsjprs.2017.03.001.
123. Wang T, Tao Y, Chen SC, Shyu ML. Multi-Task Multimodal Learning for Disaster Situation Assessment. Proceedings - 3rd International Conference on Multimedia Information Processing and Retrieval, MIPR 2020 pp. 209–212 2020. https://doi.org/10.1109/MIPR49039.2020.00050.
124. Wang Z, Ye X. Social media analytics for natural disaster management. Int J Geogr Inf Sci. 2018;32(1):49–72. https://doi.org/10.1080/13658816.2017.1367003.
125. Watson C, Cooper N, Palacio DN, Moran KP, Poshyvanyk D. A systematic literature review on the use of deep learning in Software Engineering Research. arXiv (2020).
126. Westerlund M. The emergence of deepfake technology: a review. Technol Innov Manag Rev. 2019;**9**(11).
127. Yang T, Xie J, Li G, Mou N, Li Z, Tian C, Zhao J. Social media big data mining and spatiotemporal analysis on public emotions for disaster mitigation. ISPRS Int. J. Geo Inf. 2019;(8)1:29. https://doi.org/10.3390/ijgi8010029.
128. Yigitcanlar T, Kamruzzaman M, Foth M, Sabatini J, da Costa E, Ioppolo G. Can cities become smart without being sustainable? a systematic review of the literature. Sustain Cities Soc. 2018. https://doi.org/10.1016/j.scs.2018.11.033.
129. Zhang QX, Lin GH, Zhang YM, Xu G, Wang JJ. Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images. Proc Eng. 2018;211:441–6. https://doi.org/10.1016/j.proeng.2017.12.034.
130. Zhao F, Zhang C. Building damage evaluation from satellite imagery using deep learning. In: Proceedings–2020 IEEE 21st International Conference on information reuse and integration for data science, IRI 2020 pp. 82–89 (2020). https://doi.org/10.1109/IRI49571.2020.00020.