Towards Fairness-Aware Disaster Informatics: An Interdisciplinary Perspective

YANG YANG1, CHENG ZHANG2, CHAO FAN3, ALI MOSTAFAVI3, AND XIA HU1, (Member, IEEE)
1Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77840, USA
2Department of Construction Science and Organizational Leadership, Purdue University Northwest, Hammond, IN 46323, USA
3Zachary Department of Civil Engineering, Texas A&M University, College Station, TX 77840, USA
Corresponding author: Cheng Zhang (zhan4168@pnw.edu)

This work was supported in part by the National Science Foundation under Grant IIS-1759537 and Grant CMMI-1846069 (CAREER), in part by the Amazon Web Services (AWS) Machine Learning Award, in part by the National Academies' Gulf Research Program Early-Career Research Fellowship, and in part by the Texas A&M University Open Access to Knowledge Fund (OAKFund).

ABSTRACT Collection of information from crowdsourced and traditional sensing techniques during a disaster offers opportunities to exploit this new data source to enhance situational awareness, relief, and rescue coordination, and impact assessment. The evolution of disaster/crisis informatics affords the capability to process multi-modal data and to implement analytics in support of disaster management tasks. Little is known, however, about fairness in disaster informatics and the extent to which this issue affects disaster response. Often ignored is whether existing data analytics approaches reflect the impact of communities with equality, especially the underserved communities (i.e., minorities, the elderly, and the poor). We argue that disaster informatics has not systematically identified fairness issues, and such gaps may cause issues in decision making for and coordination of disaster response and relief. Furthermore, the isolating siloed nature of the domains of fairness, machine learning, and disaster informatics prevents interchange between these pursuits. This paper bridges the knowledge gap by evaluating potential fairness issues in disaster informatics tasks based on existing disaster informatics approaches and fairness assessment criteria. Specifically, we identify potential fairness issues in disaster event detection and impact assessment tasks. We review existing approaches that address potential fairness issues by modifying the data, analytics, and outputs. Finally, this paper proposes an overarching fairness-aware disaster informatics framework to structure the workflow of mitigating fairness issues. This paper not only unveils both the ignored and essential aspects of fairness issues in disaster informatics approaches but also bridges the silos which prevent the understanding of fairness between disaster informatics researchers and machine-learning researchers.

INDEX TERMS Crisis informatics, disasters, equitable AI, fairness, social media.

I. INTRODUCTION
Suitable communications and information management techniques are essential to inform efficient and effective decision making, operational coordination, and public information and warning before, during, and after disasters. The gap between the high necessity and low availability of information magnifies disasters’ threat to human life and property [1]. Effective disaster management depends upon accurate and timely communications via dependable information systems to provide a common operating picture to government/non-governmental organizations and individuals involved in disaster management and response processes [2]. The common operating picture denotes the system that collects, synthesizes, and disseminates disaster situational awareness information to appropriate disaster management personnel, including decision-makers, first responders, and the public [3]–[5]. In this study, disaster informatics techniques refer to the data-driven analytics approaches that improve situational awareness, decision making, and operational coordination in disaster management [6]–[8]. Examples of disaster informatics approaches include, but are not limited to, evaluating disaster impact based on social media posts and remote sensing for disaster damage assessment.
In recent decades, the disaster informatics field has expanded within both the disaster management and data analytics domains (including the areas of big data, machine learning, remote sensing, and natural language processing). One promising application is the reliance on social media as a data source for assessing disasters’ damage and for planning rescue and relief operations. Zhang et al. [9] proposed summarizing, state-of-the-art social media data analytics approaches for disaster response and relief. First, for information retrieval techniques, machine learning-based classifiers can identify only social media posts related to the disasters and classify the posts into fine-grained categories, such as transportation, utility and supply, and rescue, according to their content. Similar approaches also exist in identifying disaster-related images [10]. Sentiment analysis discerns the polarity of sentiments and emotion signals as indicators of disaster impacts on public well-being [11]. Second, for information integration approaches, state-of-the-art natural language processing approaches support the detection and characterization of disaster situations and sub-events by integrating social media posts with similar contents [12]. Furthermore, information integration also includes assisting aid-seeking and aid-providing activities. Third, information interpretation integrates the information extracted from text and images in social media posts to assess the extent of the disaster (e.g., inundation of floods and intensity of earthquakes) [13].

Remote sensing data analytics, comprising change detection, object detection and analysis, and data mining and visualization, uses imagery data from remote sensing and ground-based spatial datasets to support the assessment of physical damage, sensing of the distress of populations, and provision of actionable information. Change detection enables identifying changes on the Earth’s surface acquired in the same geographical area at different stages of a disaster. Conventional change detection approaches include image differencing, change vector analysis, vegetation index differencing, and pixel-based change detection. Object detection and analysis identify and analyze objects of interest in high-resolution remotely sensed imagery. For example, Pi et al. [14] produced convolutional neural networks (CNN) to identify critical ground assets, including damaged and undamaged buildings, vehicles, vegetation, debris, and flooded areas. Data mining and visualization is the enhancement of information gleaned from remotely sensed imagery. Disaster informatics approaches significantly improve the reliability and timeliness in decision making during disaster response and relief by harnessing the power of large amounts of data generated by advanced sensing technologies [15].

The fairness issue of disaster informatics remains under-explored despite advancements in approaches to disaster informatics. Most existing disaster informatics approaches focus on improving performance for the intended tasks (such as event detection, sentiment analysis, and topic modeling) in true-positive and true-negative rates (i.e., precision and recall). Often ignored is whether these approaches can reflect the impact on underserved communities (minorities, the elderly, and the poor) with equality. As a motivating example, some recent studies raised questions about the dataset’s representativeness from social media platforms. Oktay et al. [16] analyzed US Twitter users’ demographics and found that while the user group is diverse, problems of underrepresentation exist for some demographic groups. Users in different demographic groups tend to have different usage patterns. In a study about geographical and social disparities of Twitter usage during Hurricane Harvey [17], Zou et al. questioned whether social media data represents residents of impacted areas and whether results from analysis based on such data include all members of the impacted population. Compared to the significant influences enjoyed by traditional media as a source of information for situational awareness, the fairness of the AI-based recommendation systems they employ (e.g., information push approaches) raises another concern. Zhu et al. [18] and Chakraborty et al. [19] pointed out equality issues in the existing recommendation systems, including those commonly used for news sites. Likewise, fairness issues in media coverage of disasters also exist. Franks [20] discussed the disparity issues present in Western media coverage by comparing Hurricane Katrina, Hurricane Stanley, and other disasters.

Fairness issues have attracted researchers’ attention across policy-making, education, and machine learning for several decades [21]. This paper’s central argument is that study of disaster informatics has not systematically identified fairness issues, and such gaps may cause decision-making and coordination issues in disaster response and relief. Furthermore, the isolating siloed nature of fairness, machine learning, and disaster informatics exacerbate this gap. Informed by existing studies, this paper shows the potential fairness issues based on the well-accepted theoretical framework of fairness. This paper proposes a practical framework that mitigates fairness issues for data analytics tasks for disaster situational awareness. The proposed analytics framework allows disaster informatics researchers to become aware of potential fairness issues in existing works and identify directions for addressing these fairness issues. Section II provides a review of the well-accepted fairness evaluation criteria in the machine-learning and data mining domains to provide a theoretical background. Section III provides a comprehensive discussion of the potential fairness issues in disaster informatics tasks for situational awareness. Section IV addresses the identified fairness issue by summarizing the commonly used discrimination-mitigation approaches that can be used in the pre-processing, in-processing, and post-processing stage. Section V presents the proposed fair disaster informatics analysis pipeline that could integrate the fairness evaluation criteria, and fairness issue mitigation approaches discussed above. Section VI presents a case study utilizing the framework. Finally, Section VII concludes the paper by discussing the outcomes, contributions, limitations, and future work.
TABLE 1. Sensitive attributes identified by the Fair Housing Act and the Equal Credit Opportunity Act, from [25], [26].

| Attribute       | Protected by FHA | Protected by ECOA |
|-----------------|------------------|-------------------|
| Race            | Yes              | Yes               |
| Color           | Yes              | Yes               |
| Sex             | Yes              | Yes               |
| National origin | Yes              | Yes               |
| Religion        | Yes              | Yes               |
| Familial status | Yes              | Yes               |
| Disability      | Yes              | Yes               |
| Marital status  | Yes              | Yes               |
| Age             | Yes              | Yes               |
| Receipt of public assistance | Yes | Yes |
| Exercised rights under the Consumer Credit Protection Act | Yes | Yes |
| Race            | Yes              | Yes               |
| Color           | Yes              | Yes               |
| Sex             | Yes              | Yes               |
| National origin | Yes              | Yes               |

II. BACKGROUND: FAIRNESS EVALUATION CRITERIA

This section introduces and categorizes fairness criteria that could be relevant to the field of disaster informatics. An early researcher in the area, Guion summarized his view on fairness in 1966 as “people with equal probabilities of success on the job having equal probabilities of being hired for the job” [22], which is as an early definition of individual fairness. Other researchers focus on fairness issues of different subgroups of the population [23], [24]. Following these early researchers’ perspectives, one way to categorize modern fairness is by whether fairness focuses on groups or individuals. We first discuss group fairness criteria (II.A), then individual fairness criteria (II.B), and finally causality-based fairness criteria (II.C) that are slightly different from the previous observation-based criteria.

A fairness evaluation problem can be formulated using variables. Like most other analytics tasks, disaster informatics tasks often use a series of input and output attributes. In a general fairness evaluation formulation of disaster informatics, we define variable $X$ as attributes/features (wherein $A$ is the sensitive attribute), $Y$ as the ground truth labels of $X$, and variable $R$ as prediction result based on $X$. Combining all variables above, we define disaster informatics analysis tasks as Task ($X$, $A$, $Y$, $R$). A common way to evaluate fairness in analytics is the extent to which sensitive attributes influence analytics processes. Sensitive attributes influencing social vulnerability information, such as race, gender, age, and income, should be included in the analysis. Table 1 lists sensitive attributes required to be protected by the Fair Housing Act (FHA) and the Equal Credit Opportunity Act (ECOA). The Social Vulnerability Index (SVI) dataset by the Centers for Disease Control and Prevention provides demographic and socioeconomic status information, such as per capita income, minority population, and senior population, which can serve as sensitive attributes for studies at the neighborhood level. These sensitive attributes are relevant to fairness evaluation in disaster informatics tasks discussed in Section III.

A. GROUP FAIRNESS

Group fairness focuses on identifying whether a task’s outcome is identical for groups with different sensitive attributes. In a recent study, Barocas et al. [27] provided three fairness evaluation criteria (statistical parity, separation, and sufficiency), which are representative of metrics that focus on achieving fairness between groups of people, characterized by sensitive attributes. Statistical parity focuses on avoiding the influence of demographic group membership on predictions, while separation and sufficiency focus on equalizing the quality of prediction across different demographic groups.

1) STATISTICAL PARITY (INDEPENDENCE)

To contextualize statistical parity (independence) in disaster informatics tasks, consider a system that predicts a neighborhood’s risk of being severely impacted by an upcoming disaster. Figure 1 shows the system’s confusion matrix, where $R$ is the prediction of whether the neighborhood has a high or low risk of impact, and $Y$ is whether the disaster actually impacts the neighbor. Statistical parity (also referred to as independence or demographic parity in other studies [27], [28]) focuses on ensuring that the probability of predicting high or low risk is equal across all neighborhoods (area denoted by purple dash-dot lines in Figure 1) regardless of their sensitive attributes (e.g., average income level, demographic composition, and race). Probabilistic expression of statistical parity is

$$P(R = r | A = a) = P(R = r | A = b)$$

for all $r$ in $R$, and all combinations of sensitive attributes values $a$ and $b$. For example, an analysis fails to maintain statistical parity based on demographic groups A and B if the outcome of analysis shows more positive prediction in...
examples of demographic group A and more negative prediction in demographic group B.

By requiring no correlation between sensitive attributes $A$ and output $Y$, statistical parity is a relatively strict criterion of fairness. As presented in [29], conditional statistical parity is a relaxation of the statistical parity criteria. With conditional statistical parity, the classifier is allowed more space in prediction, since conditional statistical parity does not require probabilities of all prediction outcomes to be equal or within a certain threshold, but among only those with the same given condition. An example of conditional statistical parity would be the equity of risk prediction on neighborhoods under similar physical impacts, which is relatively relaxed compared to the corresponding statistical parity criterion that requires all neighborhoods to have the sample probability of risk prediction.

2) SEPARATION (EQUALIZED ODDS) AND SUFFICIENCY

In practice, predictions of models/systems are more or less correlated with sensitive attributes due to inherent societal and historical biases. For example, do we want to omit features such as income-level that can highly increase the accuracy of predicting the hypothetical risk prediction system introduced above only because it correlates with sensitive attributes? Hardt et al. [30] pointed out that strictly enforcing statistical parity could reduce predictions’ accuracy and utility. As an alternative, Oktay et al. [16] suggest fairness criteria that focus on the parity in the accuracy of prediction for members of each subgroup, rather than parity in prediction outcome. Separation (equalized odds in [19]) and sufficiency are two criteria that focus on equality of different measurements for prediction accuracy across different subgroups.

The accuracy of model predictions is usually conducted based on metrics such as false positives, true positives, and false negatives. In a disaster informatics task, false negatives could lead to ignorance of the situation for specific vulnerable sociodemographic subgroups (such as racial minorities), delay of rescue, and ultimately higher casualties. By requiring true positive rate (TPR) and false-positive rate (FPR) to be the same or similar within a certain threshold, separation criteria focus on avoiding scenarios such as a lower detection rate of potential impacts in a low-income neighborhood due to misrepresentation in data. Probabilistic expression of separation is

$$P(Y = y|R = r, A = a) = P(Y = y|R = r, A = b)$$  \hspace{1cm} (2)

for all $r$ in $R$, $y$ in $R$ and all combinations of $a$, $b$ in $A$. Equalized opportunity is a relaxation of an equalized odds by focusing only on equality of positive prediction since the positive outcome is usually assumed to be the more significant and beneficial outcome.

On the other hand, having a substantial level of false positives could also adversely impact the results of disaster informatics tasks. For example, false positives in an impact assessment task could lead to unnecessary disruptions and a waste of much-needed resources. With the probabilistic expression of

$$P(Y = y|R = r, A = a) = P(Y = y|R = r, A = b)$$  \hspace{1cm} (3)

for all $r$ in $R$, $y$ in $R$ and all combinations of $a$, $b$ in $A$, sufficiency criteria require the equality of probability of positive predictive values (precision) and negative predictive values across all subgroups of $A$. In disaster impact assessment tasks, if false positives, which mean the predicted impacts that do not happen in reality, are heavily skewed towards low-income neighborhoods, they likely to be flagged to violate the sufficiency criterion.

### B. SIMILARITY-BASED INDIVIDUAL FAIRNESS

Group fairness metrics provide a quantitative evaluation for fairness based on groups’ membership, which is denoted only by sensitive attributes. Being treated fairly at a group level, however, may not necessarily mean being treated fairly at the individual level [31]. As an alternative, Dwork, Hardt, and Pitassi proposed fairness through awareness or individual fairness in other studies [26]. The essence of the criteria proposed by Dwork, Hardt, and Pitassi can be described as “similar individuals (measured by non-sensitive attributes) should get the same treatment.” Probabilistically, the statement can be expressed as

$$|R(x_0) - R(x_1)| \leq M(x_0, x_1)$$  \hspace{1cm} (4)

$R$ is the classifier’s outcome based on input $x$, and $M$ is a chosen distance metric.

An example of such treatment for individual fairness is in the training of job interviewers by presenting them with profiles of candidates that are similar in most aspects other than certain sensitive attributes, commonly race and gender, and testing if interviewers’ judgment would be different due to discrimination related to these sensitive attributes. Similarly, in machine learning, researchers address the problem by evaluating the closeness between the distribution of prediction $Y$ and attribute $X$ distribution with a distance metric. The distance metric, however, can greatly influence the outcome of similarity-based individual fairness, and finding a specific distance metric that fits well with the analysis task can be difficult.

### C. CAUSALITY-BASED FAIRNESS

The previously discussed fairness criteria are observation-based, as they focus primarily on joint distribution between task variables, including sensitive attribute $A$, feature $X$, prediction $R$, and outcome $Y$. Causality-based metrics focus on analyzing fairness using a graph of causal probability between the attributes to determine if sensitive attributes influence the outcome of the prediction. For example, a disaster risk prediction system could use insurance premium data in predicting the risk of disaster. Although the insurance premium attribute itself may not be a sensitive attribute, there might be a causal relationship between the insurance premium and socioeconomic status of insurance holders, which
could eventually lead to bias in the final outcome of the system. Causal analysis of the relationship between attributes can uncover such a connection.

A causal notion can capture both individual fairness and group fairness [28]. Through a comparison with the concept of “similar people should get similar treatments,” Kusner et al. [32] came up with the concept of counterfactual fairness, in which fairness is defined as each individual receiving the same treatment regardless of demographic group. From a probabilistic perspective, counterfactual fairness states that the output of the model/classifier should have the same statistical distribution for every individual in both the current setting and in an alternative, a counterfactual setting where the subject is a member of another demographic group. Bonchi et al. [33] measured both group and individual discrimination through Suppes-Bayes Causal Network (SBCN), where each node in the network represents an attribute-value pair, such as gender = male or output = positive, and weights of edges from the node represent the probabilistic causal relation from the node to its neighbors. Bonchi et al. [33] used a few random walk methods on networks built from UCI datasets (adult income data, German Credit Dataset) to model types of discrimination, including group and individual fairness. To test the method, Bonchi et al. applied the discrimination metrics to standard fairness analysis datasets, such as the German Credit Dataset, Adult Income Dataset, and Berkeley Admission Dataset. Table 2 summarizes fairness assessment metrics discussed in this section.

| Category         | Fairness Metrics                                                                 | Definition                                                                                   |
|------------------|----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| Group            | Statistical parity                                                              | \( P(R = r \mid A = a) = P(R = r \mid A = b) \) for all \( r \) in \( R \) and \( a, b \) in \( A \) |
| Separation       |                                                                                   | \( P(R = r \mid Y = y, A = a) = P(R = r \mid Y = y, A = b) \) for all \( r \) in \( R \), \( y \) in \( Y \) and all \( a, b \) in \( A \) |
| Sufficiency      |                                                                                   | \( P(Y = y \mid R = r, A = a) = P(Y = y \mid R = r, A = b) \) for all \( r \) in \( R \), \( y \) in \( Y \) and all \( a, b \) in \( A \) |
| Individual       | Fairness through awareness                                                       | \(| R(x_0) - R(x_1) | \leq M(x_0, x_1) \) for all \( x_0, x_1 \) in \( X \), where \( R \) is the outcome of the classifier based on input \( x \), and \( M \) is a chosen distance metric |
| Causality        | Counterfactual fairness                                                          | \( R(x) = R(x') \) for all \( x \) in \( X \) and its counterfactual \( x' \) |

Based on the review of Zhang et al. in [9], event detection can be partitioned into two main branches: (1) detection of disaster occurrences at a large scale and (2) detection of smaller scale sub-events (such as the closure of roads and schools). Using supervised learning filters tweets containing desired content, Yin et al. [34] proposed a supervised-learning classifier to extract disaster-related tweets based on words as well as on metadata of social media posts. Sakaki et al. [35] used a combination of supervised learning and probabilistic spatiotemporal models to detect a disaster and track its temporal and spatial patterns. Sakaki et al. [35] first built a binary support vector machine classifier using text-based features in Twitter posts to filter tweets referring to a disaster. By treating each tweet as a sensor, the researchers then used probabilistic models to trace the path of disasters.

The burst detection method recognizes abnormally frequent occurrences of disaster-related keywords in social media posts over a short window of time as an indicator of an event. A component of event detection systems, it relies on social media platforms and is sensitive to the temporal and spatial distribution of tweets related to the disaster, usually using keyword matching. Burst detection approaches identify disaster events by detecting spikes of posted information of certain topics or with certain keywords.

The demographic of users posting disaster-related content, however, might be different from that of the general public. Such disparity might exist in the distribution of disaster-related and geo-referenced content, which is supported by existing studies. Using a dataset of geo-referenced tweets, Li et al. [36] showed that users with higher socioeconomic status posted a higher density of tweets across the state of California. In another study, an analysis by Zou et al. [17] of a Twitter dataset collected before, during, and after Hurricane Harvey showed that users from communities with higher socioeconomic status are more likely to post disaster-related content on Twitter. By relying on surveillance of disaster-related tweets, burst-based event detection might be more likely to ignore events in less populous areas and areas where barriers (such as lack of access to communication technologies) prevent people from posting disaster-related situations.

III. FAIRNESS ISSUES IN DISASTER INFORMATICS

Having reviewed the general fairness issues in data mining and machine learning, in this section, we examine possible fairness issues in disaster informatics tasks. In particular, we focus on two categories of disaster informatics tasks: event detection and impact assessment.

A. FAIRNESS ISSUES IN IDENTIFYING DISASTER EVENTS

1) EVENT DETECTION FROM SOCIAL MEDIA POSTS

Event detection focuses on the rapid detection and recognition of topics in social media posts from impacted areas.
Supervised learning approaches for event detection depend on both the feature used and the algorithm itself. For features extracted from text content, one often-ignored fact is the disparity of language style among social media posts across demographic groups distinguished by age, gender, income level, and geographic location [37]. Without careful consideration, such differences can elicit unfair classification and exclusion of some tweets simply due to writing styles and languages. For example, a tweet collection is annotated with labels related/non-related to disaster as training data. Another set of tweets needs to be labeled as related/non-related. The ground truth of the tweets and the tweet labeling model are characterized by the shape and color of the Twitter icon representing each tweet data point (Figure 2), with label \( R \) being related/non-related to a disaster and signified by the blue-and-black icons. A square or oval shape indicates its ground truth \( Y \) being related/non-related. The sensitive attribute \( A \) is the socioeconomic status of the authors of these tweets (represented by the number 0 and 1 in each Twitter icon). The training dataset suffers from statistical disparity at the socioeconomic level, with the majority of disaster-related tweets coming from high-income neighborhoods, similar to the situation Zou et al. [17] suggested. As a result, classifiers trained with the biased training dataset display unfair prediction on test datasets.

Unsupervised learning or clustering is an event detection technique that assigns tweets with similar text features to a cluster. A clustering system would report a topic or event based on a threshold driven by constraints, such as the cluster’s size or the number of clusters. A common fairness issue is clustering based on text content without consideration of the variation of language style between different demographic groups, as mentioned above in supervised learning. Chierichetti et al. [38] show that the classical clustering algorithm presents significant disparity, and such disparity intensifies as the number of clusters increases. Social media posts often involve a substantial amount of noise, which could induce biases due to a large number of clusters. For instance, a filtering mechanism uses unsupervised learning to extract a cluster of event-related tweets from a set of tweets. Suppose that the majority of posts are in English. In that case, posts in other languages could be interpreted as noise by the clustering algorithm, potentially resulting in lower event detection quality in communities where English is not the most common language.

2) GEOPARSING OF EVENTS

Locating an event can significantly improve the efficiency of disaster response and resource allocation. Geoparsing, sometimes called geocoding, is mapping a location name mentioned in the text to precise geolocation. Geoparsing is an important component of disaster situational awareness tools [34]. In disaster informatics, geoparsing associates social media posts to locations to conduct related spatial analysis. The overall process of geoparsing involves the identification of location phrases in social media posts and searching for the most probable solution using internal/external knowledge bases.

Named-entity recognition (NER) methods extract name entities from text, including, but not limited to, the names of individuals, organizations, locations, facilities. Bias, however, may present in some NER systems as well. Mehrabi et al. [39] tested state-of-the-art NER systems and found a disparity in entity classification among them. Using a dataset generated from sentence templates, the researchers showed that female names are more likely to be misclassified as location entities than male names, violating both statistical parity and separation. Missing values in geoparsing systems may also be geography dependent and sociodemographic dependent. In a medical study, Oliver et al. [40] used the geoparsing function in ArcGIS with a StreetMap USA 2000 database on a dataset of prostate cancer patients. The researchers found that the percentage of addresses unable to be geocoded is associated with the percentage of elder populations and the percentage of the undereducated population, especially in rural counties.

Biases in geoparsing methods demonstrated by these studies raised some concerns about fairness issues in NER systems and pre-existing geographical databases, which could impact the outcome of analyses dependent on geoparsing data. A system to detect sub-events during disasters in an area depicted in Figure 3A might suffer from either poor NER accuracy in certain lower-income areas (Figure 3B) and/or absence of information in its geographical database in

![FIGURE 2. Illustration of results of a supervised learning task under fair classification and unfair classification.](image-url)
lower-income communities. In either case, it would result in lower accuracy in identifying location entities in lower-income areas (Figure 3C), which violates the separation and sufficiency criteria shown in section II.A.2.

**B. FAIRNESS ISSUES IN DISASTER IMPACT ASSESSMENT**

Other than detecting the occurrence of a disaster, assessment of disaster impacts (both social and physical [41]) is another essential task in disaster informatics. This section discusses potential fairness issues in conventional techniques (such as sentiment analysis) of analyzing impacts of a disaster.

1) **SENTIMENT-BASED SOCIAL IMPACT ANALYSIS**

Social impact analysis aims to understand better the influence of disaster-induced community disruptions on the impacted population. Mass media news sources, as well as crowd-sourced information over social media, allow large-scale sensing of local situations, including social impacts. Backfried et al. [42] presented a crisis management model for combining traditional media data with social media data. Sentiment analysis, in its simplest form of quantifying positive and negative sentiments in a piece of text, can be applied to aggregated or individual social media posts, as well as to news articles. Dong et al. [43] used positive/negative sentiment in tweets correlated to Hurricane Sandy to indicate the users’ general emotion in the wake of Hurricane Sandy’s arrival. Similarly, Bao-Khanh and Vo [44] tracked the general public’s emotion changes during multiple earthquakes. Others used sentiment scores as a standard feature for further analysis. Schulz et al. [45] developed classifiers using sentiment features to sort social media posts into several emotion categories. Dong et al. [43] used the structured format of news articles to calculate sentiment scores from authors, readers, and text perspectives.

Although often used in disaster informatics, pre-trained sentiment analysis systems could reveal fairness issues. Kiritchenko and Mohammad [46] analyzed more than 200 sentiment analysis systems submitted for SemEval 2018. They found that existing systems can be prone to biases, either from pre-trained embedding or the model itself. Using a set of specific sentence templates, the researchers generated a controlled dataset of sentences for test purposes, the equity evaluation corpus (EEC). The main differences between the test sentences are emotion labels and sensitive attributes. The emotion of the sentence is indicated by a defined set of words, such as happy, heartbreaking, and sadness. Sensitive attributes in the sentences are people’s names, composed of common first names of different demographic groups: African-American and European-American males and females annotated by research participants. Kiritchenko and Mohammad tested each system using the EEC dataset. Using mean score difference from paired t-test of the prediction outcome, Kiritchenko and Mohammad found a statistical difference between the prediction of different demographic groups in some systems, especially for systems that performed well on the dataset.

The results suggested that bias is potentially related to the composition of training data, pre-existing resources used (embeddings, lexicons), and the algorithms used. Such bias is an apparent violation of statistical parity criteria, where sensitive attributes, such as names of subjects, should not influence social impact prediction through sentiment scores.

A common approach for sentiment analysis is using word embedding. Word-embedding methods are models that vectorize words or sentences, which can reflect the similarity between text pieces by calculating the angle between their corresponding vectors. Word embedding uses features from existing text documents, such as Wikipedia and Google News, which may carry historical and social biases. Extracted by either traditional machine-learning models, such as support vector machine, or neural-network-based models, such as long-short-term-memory, word embedding can capture the historical and social bias presented in the documents. Diaz et al. [47] tested 15 sentiment analysis systems as well as ten widely used GloVe word-embedding models and found significant age bias in both sentiment analysis systems and embedding models. The study by Diaz et al. [47] shows that pre-trained embedding could be a possible source of
bias for sentiment analysis systems. Zhao et al. [48] applied tests on state-of-the-art sentence encoders like ELMo and BERT and found that bias also exists in these systems. In another study, Caliskan et al. [49] demonstrated that GloVe and word2vec, two common word-embedding methods, presented with human-like bias. Such bias in embedding systems could eventually influence the final result of social impacts analyses.

We argue that biases may exist if some systems proven to be biased are used in disaster informatics-related studies. For example, consider a damage assessment model that uses a sentiment score of the statement about an area or a disruptive event as an essential feature, the generation process of which involves a pre-trained embedding model and might induce bias in the process of text embedding. Figure 4 demonstrates the process of how such a biased embedding system could influence the final result of damage assessment. Given a message, the pre-trained model displays bias on word embedding of neighborhood names, which is supposed to be neutral. Being dependent on embedding, sentiment analysis systems return inconsistent results in the different neighborhoods due to non-neutral word embedding of neighborhood names. Similar to the biased sentiment analysis, impact assessment models are also influenced by the biases rooted in the biased embedding of neighborhood names.

2) PHYSICAL IMPACT ANALYSIS
Physical impact analysis, which facilitates an improved understanding of the damage in the built environment at a disaster scene, promotes better distribution of rescue resources and could reduce casualties. The standard data sources for physical impact include remote sensing data, hydrology data, and utility data. Coalescing of flooding data is also an aspect of physical impact analysis, but with many approaches. Albuquerque et al. [50] combined social media data with other data sources, including water level and digital elevation, to analyze flooding levels. Another approach uses computer vision techniques to discern flooding information. Both Liu and Wu [51] and Amit et al. [52] used CNN, a standard model for computer vision tasks, to locate landslides from satellite images. Images of disaster scenes from other sources, such as social media and news media, can be analyzed to provide physical damage inventory. Bai and Yu [6] used Mask-RCNN to detect flood depth by calculating water depth from the estimated height from photos of persons standing in water using critical attributes of the person extracted by neural network models. Alam et al. [53] suggested an image management and classification system for use in disaster rescue, with a pre-processing module removing irrelevant or duplicate images, a crowdsourcing module to acquire image labels, and finally a neural network-based classifier to categorize images into disaster-specific categories.

Wide usage of computer-vision models is linked to fairness issues in recognition found by machine-learning researchers, especially for recognition involving human faces or body silhouettes. Du et al. [54] pointed out that computer vision methods based on deep neural networks can exhibit disparity in both outcome and quality of predictions on different groups, similar to the definition of statistical parity, separation, and sufficiency presented in Section II.A. The study by Du et al. also concluded that the disparity is due to an under-representation of certain demographic groups in image datasets often used for computer-vision training. Discrimination via representation as another form of discrimination is discussed in the paper. Due to the learning capability of deep neural network models, the models themselves can extract features that are indirectly representative of sensitive attributes from a dataset with no direct connection to sensitive attributes. Table 3 summarizes the potential fairness issues for disaster informatics tasks discussed in III.A and III.B.

C. CHOOSING FAIRNESS ASSESSMENT CRITERIA FOR DISASTER INFORMATICS TASKS
The selection of fairness evaluation and mitigation is essential and could significantly impact the outcome of the analytics pipeline. Mitchell et al. [28] described a thorough process for selecting fairness criteria for prediction-based machine learning. Extending the idea of the paper into disaster informatics, we summarize trade-offs between each criterion and when to apply them.

1) WHEN TO CHOOSE GROUP FAIRNESS AND WHICH GROUP FAIRNESS CRITERIA TO CHOOSE
Due to their simplicity, the group fairness criteria apply to most data informatics tasks and are the most widely used data analytics criteria. Assessing the group fairness of disaster informatics tasks requires only the output $Y$, the predicted value $R$, and the sensitive attributes $A$. Without losing generality, we use the tasks of detecting rescue information on social media platforms as an example. It is legitimate to assess
TABLE 3. Summary of potential fairness issues for disaster informatics tasks discussed in subsections 3.1 and 3.2.

| Category                      | Definition                                                                 |
|-------------------------------|-----------------------------------------------------------------------------|
| Event detection               | • Imbalanced distribution of social media posts, which can lead to misrepresentation  
                                |   • Filtering of social media posts by supervised/unsupervised learning can lead to unintended omission |
| Geoparsing                    | • Unequal matching accuracy of name-entity recognition (NER) in different areas  
                                |   • Lack of information in external geographical databases in certain areas |
| Social impact analysis        | • Biases in word embedding systems due to historical or social biases, which can influence the outcome of social impact analysis |
| Physical impact analysis      | • Underrepresentation of training dataset in deep neural network-based computer vision methods, which is commonly used in physical impact analysis  
                                |   • Deep neural network capable of learning unintended and indirect connection between sensitive attributes and output |

the degree to which the system fairly detects rescue requests posted by persons of varied races or ethnicities. On the other hand, assessing whether the system can detect rescue requests posted by individuals with a similar situation with an equal probability is nearly impossible due to the limitations of capturing individuals by existing data-sensing approaches.

Researchers usually assess all three aspects of group fairness (statistical parity, separation, and sufficiency) when assessing the group fairness of disaster informatics approaches. It is worth noting that usually, all three aspects cannot be satisfied at the same time. Multiple studies have shown the mutual exclusivity between group-based fairness criteria of statistical parity, separation, and sufficiency except for individual cases. Barocas et al. [27] showed that statistical parity could not hold together with separation unless output $Y$ (ground truth variable of $X$) is statistically independent with sensitive attribute $A$, or $Y$ is independent with prediction result $R$. Similarly, statistical parity and sufficiency cannot both hold unless $Y$ is independent of $A$. Barocas et al. [27], Chouldechova [55] and Kleinberg et al. [56] showed that it is impossible to achieve sufficiency while satisfying separation unless for extreme cases. Deciding the trade-off among these three group fairness criteria should depend on the project’s details, balancing between prediction outcome itself or prediction accuracy.

2) WHEN TO CHOOSE INDIVIDUAL FAIRNESS

Individual fairness is often a stricter criterion compared to group fairness [56], which should be used when group fairness cannot satisfy needs. With individual fairness, one must be aware of its conflict with statistical parity and reliance on a customized distance metric. Dwork et al. [31] showed that individual fairness might be incompatible with statistical parity when the distribution of attributes in different groups is not similar. Let us use a system determining the amount of financial aid provided to disaster victims as an example. This system may compromise the sense of fairness even if its performance achieves group fairness. In other words, individuals may suffer from being overlooked or denied financial aid when the entire community with sensitive attributes (e.g., the minorities) are provided with fair aid. When using the criterion of individual fairness, researchers should focus on generating task-specific similarity measures, which is the core step of assessing individual fairness [32].

3) WHEN TO CHOOSE CAUSALITY-BASED CRITERIA

Like individual fairness, causality-based criteria also stand out when group fairness does not reflect the rationale of fairness for certain disaster informatics tasks. Specifically, cases exist in which a disaster informatics system satisfies group fairness criteria according to the testing data. Still, the sensitive attributes are used in the machine-learning model together with other intermediate variables. In this case, researchers can assess the system’s causality-based criteria to ensure the sensitive attribute will not be involved in the decision-making process for future datasets. Barocas et al. [27] noted that the fairness criterion based on end-to-end observation of $X$, $Y$, $R$ often lacks insight on the intermediate steps, resulting in the vagueness of whether sensitive attributes are involved in these steps. Bonchi et al. [33] showed that relying only on end-to-end correlation analysis could lead to false-negative discrimination reports.

IV. DISCRIMINATION MITIGATION APPROACHES

Discrimination mitigation approaches can adjust different parts of the analytics process to solve fairness issues in disaster informatics. Multiple existing studies focus on fair machine learning frameworks (e.g., Bellamy et al. [57], Mehrabi et al. [26], and D’Alessandro et al. [58]) follow the structure of pre-processing (adjust dataset), in-processing (adjust model), post-processing (adjust outputs). To assist disaster informatics researchers, we introduce a collection of mitigation methods that focus on different analytic pipeline stages. Different mitigation methods come with different constraints. Like evaluation methods, the selection of mitigation methods should be dependent on the specific disaster informatics tasks. This section will illustrate the discrimination mitigation approaches based on a hypothetic example stated in Section III.A.1: Event detection based on social media posts.
For supervised learning tasks, adjusting dataset labels is a common approach to modifying datasets to meet fairness criteria. Using binary classification as an example (e.g., labeling social media posts as related to the disaster or not), Kamiran and Calders [59] presented a method of flipping labels of an equal number of points from each protected group to make the distribution of examples between subgroups more balanced. This approach used a ranking algorithm that prioritizes examples closer to the decision boundary (the split between positive/negative examples in feature vector space). Following the notation of situational testing, Luong et al. [60] suggested a negative datapoint of the protected group to be discriminatory if its label is different from other data points with similar non-sensitive attributes. By changing the negative discriminative data points to positive, the resulting datasets show much-reduced discrimination in the adult income dataset for common classifying methods while retaining roughly the same accuracy level. Figure 5A illustrates an application of label-flipping data using the example of fair labeling of disaster-related tweets. Label-flipping approaches will change the label of the tweets in the grey area to ensure both negative and positive annotation exists for the communities with sensitive attributes.

Kamiran and Calders [59] suggested two other pre-processing techniques: reweighting and resampling. Reweighting takes a less aggressive method by modifying weights to make examples with certain attributes more significant during the training data. Figure 5B illustrates the example application of reweighting data using the example of fair labeling of disaster-related tweets. In this example, the reweighting approach will increase the weight of positive examples posted by the low-income users and the negative examples posted by users from the high-income community. Resampling examples from the original dataset could also be used to distribute examples more balanced in protected attributes. Compared to reweighting methods, resampling methods work for analytics methods that do not accept weights. Figure 5C illustrates the application of resampling data using the fair labeling of disaster-related tweets. In this example, the reweighting approach will remove the disaster-related tweets posted by the high-income communities to balance the positive training instances from the communities with or without sensitive attributes.

B. MODIFYING ANALYTICS MODELS (IN-PROCESSING)
Modifying datasets may not be possible in every situation, and finding an optimal mapping for the entire dataset while preserving the originality of the dataset can be difficult. An alternative approach for mitigating bias in the dataset is through modifying the analytics process, otherwise known as in-processing. Consider the same event detection task in section III.A, with term frequency (TF) being an important feature. Since disaster-related tweets are heavily biased towards high-income communities, without intervention, it is possible that the machine-learning model could discern the relationship between certain term-frequency features unique to tweets from high-income communities and tweets being labeled as disaster-related. Such TF features could be biased towards high-income communities and should be addressed in fairness mitigation.

Fair encoding addresses the issue by converting features to remove discriminatory associations with sensitive attributes without changing label Y. Similar to the method proposed by Calmon et al. [61], Zemel et al. [62] proposed fairness mitigation as an optimization problem of finding a fair encoding function for input feature X. To eliminate influences of protected attributes, the mapping function first needs to satisfy statistical parity constraint. While achieving statistical parity, the mapping function should lose as little information from the original dataset as possible.

Without modifying features, such as TF features related to high-income communities in the event-detection task, another in-processing approach is to penalize discrimination during model training, also known as regularization. D’Alessandro et al. [58] showed that one source of bias in the model training process is an insufficient penalty for discrimination in the model. The regularization approach is accessible since many machine-learning packages accept custom penalty functions and are flexible enough to allow researchers to reduce different types of discrimination by customizing the penalty metrics. A variety of examples of regularizers are implemented to address different fairness criteria. Methods of Kamishima et al. [63] emphasized penalizing indirect prejudice, a hidden correlation between the sensitive attributes and the target variable. Correlation between biased TF features and prediction output can be measured by indirect prejudice and can be used to penalize unfairness in our hypothetical event-detection task.

C. MODIFYING OUTPUTS (POST-PROCESSING)
In practice, re-training a fairness-aware model might also not be possible due to the constraints of time, data, and computational resources. In such scenarios, post-processing methods can be used to adjust the model’s output to be discrimination-free at a relatively low cost. Suppose that we
want to analyze disaster impact using sentiment analysis, yet we suffer from biased word-embedding systems such as those shown in section III.B.1. Post-processing methods can help to mitigate bias in embedding outputs without modifying word-embedding systems. Bolukbasi et al. [64] debiased embedding systems using a two-step process, identifying gender subspace and debiasing. Using gender bias in word2vec embedding as a case study, Bolukbasi et al. identified a list of pairs of words with gender correspondence, like (she, he), (grandma, grandpa). Using Principle Component Analysis (PCA) on the list of gender-specific word pairs, Bolukbasi et al. were able to extract a top principal component that significantly reduces variance. By assuming that this PCA captures the gender difference in the word pairs, a gender subspace is identified. Any word supposed to be gender-neutral, such as doctor and nurse, should ideally be neutral along the gender axis. Following this notation, gender bias words are re-embedded to the center on the gender axis by eliminating their component along the gender axis. The entire process does not require modification to the original embedding. It can be easily achieved through a list of calibration word pairs that can either be manually annotated or obtained from external sources.

With a similar idea of adjusting outputs by their distribution, Hardt et al. [30] presented a fairness mitigation method using only the joint distribution of \((X, A, Y)\), where \(X\) is the prediction, \(A\) is the sensitive attribute, and \(Y\) is the label. Most classifiers predict labels based on a predicted score and output labels based on the threshold of the score. Using the distribution of \((X, A, Y)\), Hardt et al. achieved equalized opportunity by setting different thresholds for each demographic group, denoted by the value of \(A\). The method by Hardt et al. can easily be applied to many existing methods. Here, we would like to illustrate the post-processing approach for mitigating the fairness issue using disaster event detection examples. The disaster-related tweet-labeling model will generate, for each tweet, a probability of relating to a disaster, and a tweet will be labeled as related to disaster if the probability exceeds a certain threshold. A standard labeler will use a uniform threshold for all tweets. A post-processing approach will identify a lower threshold for the user’s tweets from the low-income community to balance the number of disaster-related tweets posted by that and higher-income communities.

V. VISION: FAIR DISASTER MANAGEMENT ANALYSIS PIPELINE

To promote fairness-aware disaster informatics, we envision a conceptual fair disaster analysis pipeline (represented in Figure 6), inspired by comparable fair data analytics and machine learning pipelines [57], [58]. The main vision of this pipeline is to provide disaster informatics researchers with a framework for conducting fair analytics, using fairness evaluation and mitigation methods demonstrated in Section II and Section IV. Mehrabi et al. [26] discussed that fairness intervention should be domain-specific. Similarly, actual methods to be used in fairness evaluation and mitigation should be carefully chosen based on the specific disaster informatics tasks.

The pipeline is similar to a regular data analytics pipeline, but with extra fairness evaluation steps on both dataset and analytics tasks and fairness mitigation steps for results and models. The pipeline starts with extracting data and features from external sources. Then the generated dataset goes through a dataset fairness evaluation process. If the dataset fails the fairness evaluation, adjustments need to be made. We can use causality analysis to find attributes with an indirect causal relationship to protected attributes for causality-related issues. For statistical party-related issues, we can evaluate using statistical fairness metrics or similarity-based metrics if we want to address individual fairness. Pre-processing methods should be selected according to fairness criteria as well. After adjustment, the initial dataset should be checked against fairness evaluation again.

Fairness evaluation of analytic methods is similar to that of a fair machine-learning process but modified and expanded to fit into general analytics tasks. We applied the data through each disaster informatics task, depending on the specific study. Given a set of inputs and corresponding outputs from the task, researchers can evaluate most fairness metrics including statistical metrics and similarity-based metrics. If a certain output fails the fairness evaluation, it can be either applied to fairness post-processing methods or to in-processing methods. For in-processing, analytic tasks are rerun with a more optimized setting to remove discrimination, such as including regularizers based on different discrimination metrics or adding an intermediate encoding to remove discriminatory features. For post-processing, the outputs are adjusted using the desired post-processing methods, and the transformed requires modifying the original analytics task, which might not be possible when using third-party tools. Post-processing methods usually approach fairness issues by setting new demographically based thresholds for prediction scores and are usually more lightweight than in-processing methods. After the post-processing step, outputs will again go through results that can be directly outputted. In-processing analytics fairness evaluation. The loop will be iterated until the fairness criteria are satisfied. Final outcomes should be relatively fair regarding sensitive attributes in a disaster informatics context. In addition, this pipeline can be applied to the entire analysis or to any sub-process, such as a geoparsing system or sentiment analysis, where the output will be used as features in later steps.

VI. CASE STUDY

The case study revolves around assessing the tweets’ sentiment scores posted by the resident affected by Hurricane Harvey. The sentiment score can be an indicator of disaster impact. The input datasets were based on tweets collected from the Houston area during Hurricane Harvey (from August 25, 2017, through August 31, 2017). Among these tweets from Houston, more than 23,000 tweets were tagged with precise geo-coordinates, which were selected for our
FIGURE 6. Fair disaster informatics pipeline vertically divided into three significant components, data (pre-processing), analysis (in-processing) and output (post-processing).

input dataset. Sentiment scores were calculated for each tweet with Vader sentiment analysis [65]. Afterward, sentiment scores were aggregated to census tract level by taking the average of sentiment score of tweets located within each census tract. As sensitive attributes of each census tract, estimated per capita income (E_PCI), the ratio of the population age 65 or above (E_AGE65), and the ratio of minority population (E_MINRTY) were derived from the Social Vulnerability Index published by the Centers for Disease Control and Prevention. Each census tract’s sentiment score was converted to binary, representing above or below the median of all census tracts in the Houston area. Similarly, the sensitive attributes were also converted to binary, representing above or below median across the nation. With sentiment score being outcome variable $Y$ and E_PCI, E_AGE65, E_MINRTY being sensitive attribute $A$, statistical parity can be calculated for the task. In table 4, the largest difference between the probability of the above sentiment outcome is the one between census tracts with above-median E_MINRTY and below-median E_MINRTY, followed closely by E_PCI, and then by E_AGE65. We can conclude that the analysis outcome fails the fairness criterion of statistical parity.

We can determine the disparity in probabilities due to our analysis, such as biases in a word embedding against the names of individuals of certain demographic groups or names of neighborhoods with certain socioeconomic status. In that case, we should intervene and mitigate such bias. To mitigate such bias in sentiment score, we proposed two approaches, mitigation at training and mitigation at the output, which corresponds to the in-processing and post-processing approach in section IV. For in-process mitigation, Zhao et al. [66] suggested a method to re-train embedding models to reduce bias.

If the word-embedding output of a sentence would be reduced if it contained names of neighborhoods with low E_PCI, this bias can be modeled in the embedding output as the differences between sentences whose subjects are neighborhood names. A penalty for the bias can be introduced into the training process of word-embedding models to allow the model to recognize and reduce such biases. Modifying and re-training the embedding model, however, may not be a practical option for researchers not familiar with the models or with machine learning.

Alternatively, we can apply the post-processing approach if we determine a bias is due to sentiment analysis, similar to situations shown in [46], [64]. For instance, if a sentiment analysis system contains bias towards sentences containing information about lower income neighborhoods and would give them a lower sentiment score, Boluski’s method [64] can be used to mitigate the sentiment output without changing the model.

VII. CONCLUDING REMARKS

A critical blind spot in the existing disaster informatics literature is the rather limited consideration of fairness issues. In disaster informatics, approaches that can reflect different communities’ impact with equality, especially for underserved communities (minorities, the elderly, and the poor) are often ignored. To promote fairness in disaster informatics, this study argues that disaster informatics has not systematically identified fairness issues, and such gaps may cause decision making and coordination issues in disaster response.

| Sentiment Score | Percentage of minority population | Per capita income | Percentage of aged population ($\geq 65$) |
|-----------------|----------------------------------|-------------------|--------------------------------------|
| Below median    | 0.067                            | 0.011             | 0.002                                |
| Above median    | 0.147                            | 0.145             | 0.101                                |
and relief. Furthermore, the isolated nature of the domains of fairness, machine learning, and disaster informatics exacerbates such a gap. This paper bridges the knowledge gap by evaluating the potential fairness issues in disaster informatics tasks based on existing disaster informatics approaches and fairness assessment criteria. Specifically, we identified the potential fairness issues regarding disaster event detection and impact assessment tasks. We further reviewed approaches to address the potential fairness issues by modifying the data, the analytics models, and the results. Finally, this paper proposes a fairness-aware disaster informatics framework structuring the workflow of mitigating fairness issues.

This paper’s theoretical contribution is that it is the first study, to the authors’ knowledge, that identifies potential fairness issues in disaster informatics tasks, the significance of which is two-fold. First, this study can help disaster informatics researchers gain awareness of potential biases in data sources, analytical models, outcomes that may harm the well-being of the poor, the elderly, and minorities and provide a set of mitigation tools to address the fairness issues. Second, this paper also provides the researchers in the domain of machine learning an opportunity to identify real-world problems regarding fairness issues.

The practical contribution of this paper is mainly in (1) providing scenarios and explanations of how biases are introduced in common disaster analysis processes, and (2) providing a framework to evaluate and mitigate these biases. Section III covers common analysis approaches in disaster informatics and discussed how biases can be introduced into these analysis approaches, such as the issue of uneven distribution of tweets leading to reduced awareness of events in certain areas or biases hidden in pre-trained word embedding leading to biases in impact evaluation for affected areas. We examined the sources of these fairness issues, which are critical for the public, relief organizations, and first responders to gain situational awareness and impacts of disasters. Prior disaster informatics techniques and tools may not consider fairness issues, a situation that may cause inequitable situation assessment and resource allocation. Section II and Section IV provide a comprehensive review of up-to-date fairness evaluation criteria and mitigation strategies to identify and mitigate these issues. The categories of fairness metrics are group fairness, individual fairness, and causality fairness. Fairness mitigation strategies are categorized into pre-processing (mitigation on the dataset), in-processing (mitigation during analysis), and post-processing (mitigation of result outputs). In section V, we discuss how the proposed fairness framework’s pipeline would utilize these evaluation and mitigation strategies together in a relatively intuitive manner. In section VI, we provided a case study using such a framework. With the proposed fairness framework, existing models and algorithms will have great potential to craft a fair picture of the disaster situations to support resources and action prioritization. The mitigation of inequity in analytics techniques for disaster response will further influence other scenarios such as disaster mitigation planning and facility investment.

One direction for future research into fairness-aware disaster informatics would be understanding the trade-off between fairness criteria for different disaster data analytics tasks. Another trade-off to study is between fairness and performance. Existing mitigating approaches, such as data resampling, addresses fairness issues at the cost of accuracy of analytics results. To achieve fair disaster analytics with high performance, it is necessary to have high redundancy of data quality and quantity, as well as the high performance of the model. Potential fairness issues also call for packages and toolkits for rapid detection and mitigation of the biases in disaster informatics tasks. Disaster management and response researchers and practitioners without strong computer science and engineering backgrounds may have difficulty in modifying the datasets and machine learning models. Therefore, packages and toolkits mitigating fairness issues can help future researchers to identify and address fairness issues in current studies and disaster management activities to prevent harm to and to better serve vulnerable populations.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation and National Academies.

REFERENCES

[1] I. Shklovski, M. Burke, S. Kiesler, and R. Kraut, “Technology adoption and use in the aftermath of hurricane katrina in New Orleans,” Amer. Behav. Scientist, vol. 53, no. 8, pp. 1228–1246, Apr. 2010.
[2] National Incident Management System, US Dept. Homeland Secur., Washington, DC, USA, 2008.
[3] P. P. Ray, M. Mukherjee, and L. Shu, “Internet of Things for disaster management: State-of-the-art and prospects,” IEEE Access, vol. 5, pp. 18818–18835, 2017, doi: 10.1109/ACCESS.2017.2752174.
[4] Y. Zhao and W. Song, “Survey on social-aware data dissemination over mobile wireless networks,” IEEE Access, vol. 5, pp. 6049–6059, 2017, doi: 10.1109/ACCESS.2016.2709368.
[5] L. Xiao, Y. Li, G. Han, H. Dai, and H. V. Poor, “A secure mobile crowdsensing game with deep reinforcement learning,” IEEE Trans. Inf. Forensics Security, vol. 13, no. 1, pp. 35–47, Jan. 2018, doi: 10.1109/TIFS.2017.2737968.
[6] H. Bai and G. Yu, “A Weibo-based approach to disaster informatics: Incident monitor in post-disaster situation via Weibo text negative sentiment analysis,” Natural Hazards, vol. 83, no. 2, pp. 1177–1196, Sep. 2016, doi: 10.1007/s11069-016-2370-5.
[7] R. R. Murphy, “Emergency informatics: Using computing to improve disaster management,” Computer, vol. 49, no. 5, pp. 19–27, May 2016, doi: 10.1109/MC.2016.135.
[8] S. A. Shah, D. Z. Seker, M. M. Rathore, S. Hameed, S. B. Yahia, and D. Draheim, “Towards disaster resilient smart cities: Can Internet of Things and big data analytics be the game changers?,” IEEE Access, vol. 7, pp. 91885–91903, 2019, doi: 10.1109/ACCESS.2019.2928233.
[9] C. Zhang, C. Fan, W. Yao, X. Hu, and A. Mostafavi, “Social media for intelligent public information and warning in disasters: An interdisciplinary review,” Int. J. Inf. Manage., vol. 49, pp. 190–207, Dec. 2019, doi: 10.1016/j.ijinfomgt.2019.04.004.
[10] Z. Meng, B. Peng, and Q. Huang, “Flood depth estimation from Web images,” in Proc. 2nd ACM SIGSPATIAL Int. Workshop Adv. Resilient Intell. Cities (ARIC), 2019, pp. 37–40, doi: 10.1145/3356395.3365542.
C. Zhang, W. Yao, Y. Yang, R. Huang, and A. Mostafavi, “Semiautomated social media analytics for sensing societal impacts due to community disruptions during disasters,” Comput.-Aided Civil Infrastruct. Eng., vol. 1, pp. 1–18, Jun. 2020, doi: 10.1111/mice.12576.

F. Alam, F. Oflı, and M. Imran, “Descriptive and visual summaries of disaster events using artificial intelligence techniques: Case studies of hurricanes Harvey, Irma, and Maria,” Behav. Inf. Technol., vol. 39, no. 3, pp. 288–318, Mar. 2020, doi: 10.1080/0144929X.2019.1619098.

C. Fan, F. Wu, and A. Mostafawi, “A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters,” IEEE Access, vol. 8, pp. 10478–10490, 2020, doi: 10.1109/ACCESS.2020.2965550.

Y. Pi, N. D. Nath, and A. H. Behzadan, “Convolutional neural networks for object detection in aerial imagery for disaster response and recovery,” Adv. Eng. Informat., vol. 43, Jan. 2020, Art. no. 101009, doi: 10.1016/j.aei.2019.101009.

A. Cooner, Y. Shao, and J. Campbell, “Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 Haiti earthquake,” Remote Sens., vol. 8, no. 10, p. 868, Oct. 2016, doi: 10.3390/rs8100868.

H. Oktay, A. Firat, and Z. Ertem, “Demographic breakdown of Twitter users: An analysis based on names,” in Proc. ASE BIG-DATA/SOCIALCOM/CYBERSECURITY Conf., Mar. 2014, pp. 1–11.

L. Zou, N. S. N. Lam, S. Shams, H. Cai, M. A. Meyer, S. Yang, K. Lee, S.-J. Park, and M. A. Reams, “Social and geographical disparities in Twitter use during Hurricane Harvey,” J. Dig. Earth, vol. 12, pp. 1–19, Dec. 2018, doi: 10.1080/17538947.2018.1545878.

Z. Zhu, X. Hu, and J. Caverlee, “Fairness-aware tensor-based recommendation,” in Proc. 27th ACM Int. Conf. Inf. Knowl. Manage., 2018, pp. 1153–1162.

A. Chakraborthy, G. K. Patro, N. Ganguly, K. P. Gummadi, and P. Loiseau, “Equality of voice: Towards fair representation in crowdsourced top-K recommendations,” in Proc. Conf. Fairness, Accountability, Transparency (FAIT), 2019, pp. 129–138, doi: 10.1145/3287560.3287570.

S. Franks, “The CARMA report: Western media coverage of humanitarian crises,” J. Educ. Meas., vol. 5, no. 2, pp. 20–37, Feb. 1966.

M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in Proc. 30th Int. Conf. Inf. Syst. Cris. Response Manage. (ISCRAM), May 2013, pp. 846–851.

S. Barocas, M. Hardt, and A. Narayanan, “Examining gender and race bias in two hundred sentiment analysis systems,” in Proc. 7th Joint Conf. Lexical Comput. Semantics, 2018, pp. 43–53, doi: 10.18653/v1/18-2005.

M. Díaz, J. Johnson, A. Lazar, A. M. Piper, and D. Gergle, “Addressing age-related bias in sentiment analysis,” in Proc. 28th Int. Joint Conf. Artif. Intell., Aug. 2019, pp. 6146–6150, doi: 10.24963/ijcai.2019/852.

J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K. Chen, “Gender bias in contextualized word embeddings,” 2019, arXiv:1904.03310. [Online]. Available: http://arxiv.org/abs/1904.03310.

A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” Science, vol. 356, no. 6334, pp. 183–186, Apr. 2017.

J. P. de Albuquerque, B. Herfort, A. Brenning, and A. Zipf, “A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management,” Int. J. Geograph. Inf. Sci., vol. 29, no. 4, pp. 667–689, Apr. 2015, doi: 10.1080/13658816.2014.996567.

J. Liu and L. Wu, “Geological disaster recognition on optical remote sensing images using deep learning,” Procedia Comput. Sci., vol. 91, pp. 566–575, Jan. 2016, doi: 10.1016/j.procs.2016.07.144.

S. N. K. B. A. M. K. S. Shiraiashi, T. Inoshita, and Y. Aoki, “Analysis of satellite images for disaster detection,” in Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), Jul. 2016, pp. 5189–5192.

M. Du, F. Yang, N. Zou, and X. Hu, “Fairness in deep learning: A computational perspective,” IEEE Intell. Syst., early access, Jun. 10, 2020, doi: 10.1109/MIS.2020.3000681.

A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” Big Data, vol. 5, no. 2, pp. 153–163, Jun. 2017.
[56] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” 2016, arXiv:1609.05807. [Online]. Available: http://arxiv.org/abs/1609.05807

[57] R. K. E. Bellamy, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Sahu, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, K. Dey, M. Hind, S. C. Hoffman, S. Houle, K. Kannan, P. Lohia, J. Martino, and S. Mehta, “AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,” IBM J. Res. Develop., vol. 63, nos. 4–5, pp. 4:1–4:15, Jul. 2019, doi: 10.1147/JRD.2019.2942287.

[58] B. d’Alessandro, C. O’Neil, and T. LaGatta, “Conscientious classification: A data scientist’s guide to discrimination-aware classification,” Big Data, vol. 5, no. 2, pp. 120–134, Jun. 2017, doi: 10.1089/big.2016.0048.

[59] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” Knowl. Inf. Syst., vol. 33, no. 1, pp. 1–33, 2012.

[60] B. T. Luong, S. Ruggieri, and F. Turini, “K-NN as an implementation of situation testing for discrimination discovery and prevention,” in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2011, pp. 502–510, doi: 10.1145/2020408.2020488.

[61] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, “Optimized pre-processing for discrimination prevention,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Dec. 2017, pp. 3993–4002.

[62] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in Proc. 30th Int. Conf. Mach. Learn. (ICML), 2013, vol. 28, no. 2, pp. 1362–1370.

[63] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, “Fairness-aware classifier with prejudice remover regularizer,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases, 2012, pp. 35–50.

[64] T. Bolukbasi, K. W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2016, pp. 4356–4364.

[65] C. J. Hutto and E. E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” in Proc. 8th Int. Conf. Weblogs Social Media (ICWSM), 2014, p. 82.

[66] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, “Learning gender-neutral word embeddings,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2018, pp. 4847–4853.

YANG YANG was born in Guangdong, China, in 1995. He received the B.S. degree in computer science from Texas A&M University, College Station, TX, USA, in 2019, where he is currently pursuing the M.S. degree in computer science. After joining the Urban Resilience AI Laboratory, in May 2018, he worked on multiple interdisciplinary studies of applying data analytics and machine learning methods on social media analysis. His current research interests include fairness in artificial intelligence and data analysis on social media user behaviors.

CHENG ZHANG received the bachelor’s degree in thermos energy and power engineering from Shanghai Jiao Tong University, the master’s degree in navigation, guidance, and control from the Nanjing University of Aeronautics and Astronautics, and the Ph.D. degree in civil, environmental, and sustainable engineering from Arizona State University, in July 2017. He is currently an Assistant Professor with the Department of Construction Science and Organizational Leadership, Purdue University Northwest. His research interests include automation, artificial intelligence, data analytics techniques for improving infrastructure system resilience in the face of extreme events, such as crisis, turnarounds, or natural hazards, advancing the integrated human–machine intelligence in sensing, modeling, and information management for reliable decision-making and coordination.

CHAO FAN received the Master of Science degree in civil engineering from the University of California, Davis, CA, USA, in 2017. He is currently pursuing the Ph.D. degree with the Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, USA. He joined the Urban Resilience AI Laboratory in Fall 2017. His research interests include cutting-edge interdisciplinary research at the interface of engineering, science, policy, integrated and resilience studies in disaster management using social media, and data analytics methods. He received the Doctoral Fellow and the Best Paper Nomination from the HAWAII International Conference on System Sciences.

ALI MOSTAFAVI received the Master of Science degree in industrial administration degree from the Krannert School of Management and the Ph.D. degree in civil engineering from Purdue University, West Lafayette, IN, USA, in 2013. He is currently an Associate Professor with the Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, USA. He is also the Director of the Urban Resilience AI Laboratory. His research interests include creating transformative solutions for addressing the grand challenges pertaining to the nexus of humans, disasters, and infrastructure systems.

XIA HU (Member, IEEE) received the Bachelor of Science and Master of Science degrees from Beihang University, and the Ph.D. degree from Arizona State University under the supervision of Dr. H. Liu. He was a Postdoctoral Researcher with Arizona State University and the Phoenix Veteran Affairs Health Care System, as a Research Intern in microsoft research. He and Dr. T-Seng Chua was a Visiting Student with the National University of Singapore. He is currently an Associate Professor with the Department of Computer Science and Engineering, Texas A&M University.

Y. Yang et al.: Towards Fairness-Aware Disaster Informatics