2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

# Coordinating Disaster Emergency Response with Heuristic Reinforcement Learning

Zhou Yang*, Long Nguyen†, Jiazhen Zhu‡, Zhenhe Pan†, Jia Li§, and Fang Jin*
* Department of Statistics, George Washington University
† Department of Computer Science, Texas Tech University
‡ Department of Civil and Environmental Engineering, University of California, Davis
§ Walmart Global Tech
Email: zhou_yang@gwmail.gwu.edu, {long.nguyen, zhenpan, jia.li}@ttu.edu,
jiazhen.zhu@walmart.com, cejli@ucdavis.edu, fangjin@gwu.edu

***Abstract*—A crucial and time-sensitive task when any disaster occurs is to rescue victims and distribute resources to the right groups and locations. This task is challenging in populated urban areas, due to a huge burst of help requests made in a very short period. To improve the efficiency of the emergency response in the immediate aftermath of a disaster, we propose a heuristic multi-agent reinforcement learning scheduling algorithm, named as ResQ, which can effectively schedule a rapid deployment of volunteers to rescue victims in dynamic settings. The core concept is to quickly identify victims and volunteers from social network data and then schedule rescue parties with an adaptive learning algorithm. This framework performs two key functions: 1) identify trapped victims and volunteers, and 2) optimize the volunteers' rescue strategy in a complex time-sensitive environment. The proposed ResQ algorithm can speed up the training processes through a heuristic function which reduces the state-action space by identifying a set of particular actions over others. Experimental results showed that the proposed heuristic multi-agent reinforcement learning based scheduling outperforms several state-of-art methods, in terms of both reward rate and response times.**

## I. Introduction

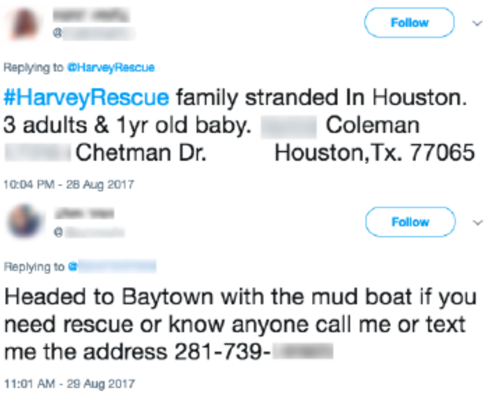
Natural disasters have always posed a critical threat to human beings, often being accompanied by a major loss of life and property damage. In recent years, we have witnessed more frequent and intense natural disasters all over the world. In 2017 alone, there were multiple devastating natural disasters, each resulting in hundreds of deaths. Hurricanes, flooding, tornadoes, earthquakes and wildfires, were all active keywords in 2017. To mitigate the impacts of disasters, it is important to rapidly match the available rescue resources with disaster victims who need help in the most efficient way, in order to maximize the impact of the rescue effort with limited resources. A key challenge in disaster rescues is to balance the requests for help with the volunteers available to meet that demand.

The adverse impacts of a disaster can be substantially mitigated if during the disaster accurate information regarding the available volunteers can be gathered and victims' locations can be determined in a timely manner, enabling a well-coordinated and efficient response. This is particularly apparent whenever there is a huge burst of requests for limited public resources. For example, when Hurricane Harvey made landfall on August 25, 2017, flooding parts of Houston, the 911 service was overwhelmed by thousands of calls from victims in a very short period. Since the phone line resource is limited, many phone calls did not get through and victims turned to social media to plead for help, posting requests with their addresses. At the same time, many willing volunteers seeking to offer help during the disaster were left idle as no one knew where they should be sent. This case is illustrated in Figure 1, along with a sample distribution of victims and volunteers in Figure 2. In the case of a hurricane, a major challenge is that without coordination, multiple volunteers with boats may go to rescue the same victim while other victims have to wait for extended times to be rescued. This mismatch between victims and volunteers represents an enormous waste of limited volunteer resources. It is therefore imperative to improve the emergency services' coordination to enable them to efficiently share information, coordinate rescue efforts and allocate resources more effectively, and offer guidance for optimal resource allocation.

The problem of resource coordination has drawn considerable attention in the computer science community, and several data mining frameworks have been developed to address this problem. Previous researchers have primarily focused on three approaches: supervised learning, adaptive methods, and optimization-based method. Traditional supervised learning models demand a dataset that is statistically large in order to train a reliable model [1], [2], for example, by building regression models to predict needs and schedule resources accordingly [3]. Unfortunately, due to the unique nature of resource management for disaster relief, it is generally impractical to model this using traditional supervised learning models. Every disaster is unique and hence it makes no sense to model one disaster relief problem by using the dataset collected from other disasters; a realistic dataset for that disaster can only be obtained when it occurs. This means that traditional supervised learning is unable to solve the highly individual resource management problems associated with disaster relief efforts.

Other researchers have developed adaptive methods [4], [5] and proposed adaptive systems [6] for resource allocation.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE



[redacted]
@[redacted]

Replying to @HarveyRescue

#HarveyRescue family stranded In Houston.
3 adults & 1yr old baby. [redacted] Coleman
[redacted] Chetman Dr. Houston,Tx. 77065

10:04 PM - 28 Aug 2017

[redacted]
@[redacted]

Replying to @[redacted]

Headed to Baytown with the mud boat if you need rescue or know anyone call me or text me the address 281-739-[redacted]

11:01 AM - 29 Aug 2017

Fig. 1: Sample tweets requesting for rescue and offering help.

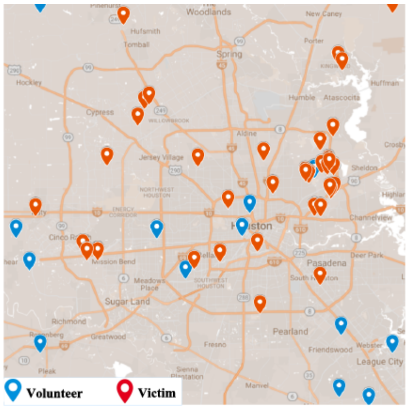


Fig. 2: The distribution of volunteers and victims in the Houston area on August 28, 2017.

However, a common limitation of the adaptive approach is that the parameters in adaptive models change slowly and hence converge slowly. An alternative is to model resource coordination problems as simulation problems or optimization problems which requires the process of modelling and tuning repeatedly if any of the external environmental parameters change.

The recent success achieved in applying machine learning to challenging decision-making domains [7]–[9] suggests that Reinforcement Learning (RL) is a promising method with considerable potential.

In this paper, we aim to find an effective way to coordinate the efforts of volunteers and enable them to reach disaster victims as soon as possible. We have developed a novel heuristic multi-agent reinforcement learning-based framework to analyze the tweets and identify volunteers and victims, along with their locations. Based on the information collected, a resource coordination system can then allocate the volunteer resources more efficiently. The resource coordination system is implemented using heuristic multi-agent reinforcement learning since this approach offers a good way to address the above dilemmas because of its unique characteristics. More specifically:

- We build an efficient heuristic multi-agent reinforcement learning framework for large-scale disaster rescue work based on information gathered by mining social media data. This study is one of the first that specifically focuses on coordinating volunteers in disaster relief using reinforcement learning.
- We propose a ResQ algorithm, which is capable of adapting dynamically as information comes in about volunteers and victims' situations and makes recommendations to minimize the total distance travelled by all the volunteers to rescue the maximum possible number of victims.
- Our proposed new disaster relief framework bridges the gap when traditional emergency helplines such as 911 are overwhelmed, thus benefiting both the disaster victims and the non-Governmental organizations seeking to help them.
- Last but not least, our proposed ResQ algorithm significantly outperforms existing state-of-the-art methods, reducing the computation times required considerably. The effectiveness of the proposed method is validated using a Hurricane Harvey related social media dataset collected in August 2017 for the Houston area, Texas.

## II. Related Work

### A. Disaster Relief with Social Media.

The most recent survey [10] pointed out that, the success of a disaster relief and response process relies on timely and accurate information regarding the status of the disaster, the surrounding environment, and the affected people. There are a large number of studies using social media data for disaster relief [11]. Gao et al. [12] built a crowdsourcing platform to provide emergency services during the 2010 Haiti earthquake, such as handling food requests. They integrated the system with crisis maps to help organizations to identify the location where supplies are most needed. [13] compared the use of social media and news in disaster event. [14], [15] forecasted human needs to enhance preparedness in disaster response. [16] proposed a normalization technique to enhance reasoning over social media content. Zook et al. [17] demonstrated that information technologies were the key means through which individuals could contribute to relief efforts without being physically present in Haiti. This example proved how to make full use of volunteer resources by outsourcing tasks to remote volunteers. Ashktorab et al. [18] introduced a Twitter-mining tool to extract practical information for disaster relief workers during natural disasters.

### B. Multi-agent Reinforcement Learning

The research on Multi-agent Reinforcement Learning (MARL) has proved to be very challenging. The exponential growth of the discrete state-action space gives rise to a challenge for iterating over the state-action space. The correlated returns of multiple agents make it difficult to maximize the returns independently. Several MARL goals have been proposed to circumvent this problem. Hu and Wellman proposed a framework where agents maintain Q-functions over joint actions and perform updates based on agents' learning dynamics [19]. Powers and Shoham proposed to consider the adaption to the changing behaviors of the other agents [20]. Other researchers also proposed to consider both stability and adaption at the same time [21]–[23].

## III. Problem Formulation

### A. Problem Formulation

*Definition 3.1 (Rescue Scheduling Task):* Let $A_t$ denote the set of assignments of victims to be rescued by volunteers at time $t$. Given a set of volunteers $U_t = \{u_1, u_2, ...u_{M_t}\}$, and a set of victims $V_t = \{v_1, v_2, ...v_{N_t}\}$, a rescue scheduling task is to find a set of sequential assignments of volunteers to rescue victims, such that all the victims are rescued with minimal total cost. The total cost for such scheduling is the total time spent on rescuing all the victims.

For the purpose of this study, we assume that victims are taken to the nearest shelter after they have been rescued. We calculate the total time for a rescue task as $T = T(D)_{travel} + T_{load} + T_{shelter}$, where $D$ is the distance between the volunteer and the victim, $T(D)_{travel}$ is the travel time that it takes for the volunteer to reach the victim, $T_{load}$ is the time to load the victim(s) to the boat, and $T_{shelter}$ is the time needed to carry them to the nearest shelter. Since the loading time $T_{load}$ and the time to shelter $T_{shelter}$ are constants in every scheduling policy, we will not take the loading time $T_{load}$ and the time to shelter $T_{shelter}$ into consideration.

Assignment $X_t \in A_t$ may be written as an $N_t \times M_t$ matrix, in which column $i$ lists the victims that volunteer $U_i$ will rescue at time $t$, in order. Suppose there are $N_t$ victims to be rescued by $M_t$ volunteers. We can now represent the rescue scheduling result as a matrix $X_t = (x_{ij})_{N_t M_t}$

$$x_{ij} = \begin{cases} 1 & \text{volunteer i is dispatched to rescue victim j,} \\ 0 & \text{volunteer i is not dispatched to rescue victim j.} \end{cases}$$

where $1 \leq i \leq N$, $1 \leq j \leq M$.

In this case, a volunteer rescues one victim at a time, while a victim can only be rescued by at least one volunteer. The mathematical model for the volunteer-victim problem is defined as follows:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & C = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \sum_{j=1}^{M_t} d_{ij} x_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{N_t} x_{ij} \leq 1,\ j = 1, \ldots, M_t; \\ & \sum_{j=1}^{M_t} x_{ij} \geq 1,\ i = 1, \ldots, N_t; \\ & x_{ij} \in \{0, 1\}. \end{aligned}$$

where $d_{ij}$ is the distance from volunteer i to victim j.

### B. ResQ: Heuristic Multi-agent Reinforcement Learning (MARL) in Rescue Scheduling

*1) The setting of MARL:* To tackle this rescue scheduling problem, we can formulate the problem using multi-agent reinforcement learning technique [24]. The agents are volunteers who are willing to rescue disaster victims. The victims represent the rewards and the environment is the place where the disaster happened. This environment is represented as a square-grid world, and the agents move within this grid world to rescue the victims. In other words, this is a Markov game G for N agents , which is denoted by a tuple $G =< \mathrm{N, S, A, P, R}, \gamma >$, where $N$, $S$, $A$, $P$, $R$, $\gamma$ are the number of agents, sets of states, joint action space, transition probability function, reward function and discount factor respectively. These are defined as follows:

- **Agent**: We consider a volunteer with a boat to be an agent.
- **State** $s_t \in S$: A state $s_t^i$ of a volunteer $i$ at time $t$ in the rescue scheduling problem is defined as the possible grid location where he or she is located.
- **Action** $a_t \in A$**=** $A_1 \times \ldots \times A_{N_t}$: a joint action $a_t = \{a_t^i\}_1^{N_t}$ denotes the allocation strategy of all available volunteers at time $t$, where $N_t$ is the number of available agent at time $t$.
- **Transition function** $P : S \times A \rightarrow [0, 1]$: The state transition probability $p(s_{t+1}|s_t, a_t)$ gives the probability of transiting to $s_{t+1} \in S$ given a joint action $a_t \in A_i$ is taken in the current state $s_t \in S$.
- **Reward function** $R_i \in R = S \times A \rightarrow (-\infty, +\infty)$: The $i - th$ agent attempts to maximize its own expected discounted reward: $R_t = E(r_t^i + \gamma r_{t+1}^i + ...) = E(\sum_{k=0}^{\infty} \gamma^k r_{t+k}^i) = E(r_t^i + \gamma R_{t+1})$.

The goal of our disaster rescuing problem is to find the optimal policy $\pi^*$ (a sequence of actions for agents) that maximizes the total reward. The state value function $V^\pi(s)$ is introduced to evaluate the performance of different policies. $V^\pi(s)$ stands for the expected total reward with discount from current state $s$ on-wards with the policy $\pi$, which is equal to:

$$\begin{aligned} V^\pi(s) &= E_\pi(R_t|S = s_t) = E_\pi(r_t + \gamma V^\pi(s')) \\ &= r_t + \sum_{s' \in S} P^\pi(s'|s) V^\pi(s'). \end{aligned} \tag{1}$$

According to Bellman optimality equation [25], we have

$$V^\pi(s) = \max_{a \in A} \{r_t(s, a) + \sum_{s' \in S} \gamma P^\pi(s'|s, a) V^\pi(s')\}. \tag{2}$$

Since the volunteers have to explore the environment in order to find victims, they cannot observe the underlying state of the environment. We treat this as a Partially Observable Markov Decision Process (POMDP) [26]. A POMDP extends the definition of Markov Decision Process (MDP). It is defined by a set of states $S$ denoting the environment setting for all agents, a set of actions $A_1 ... A_N$ and a set of observations $O_1 ... O_N$ for each agent. The state transition function $P : S \times$

**Algorithm 1:** ResQ in Rescue Scheduling

```
let t=0, $Q^i_t$=1;
initialize $s_0$;
repeat
    Observe current state $S_t$;
    $A_t$ = HeuristicActionSelection($S_t$)
    Every volunteer execute its action $a^i_t$ in $A_t$;
    Observe $R^i_t ... R_t$ and $a^i_t ..._t$

    $Q^i_{t+1}(s, a^1...a^N) \leftarrow (1-a_t)Q^i(s, a^1...a^N) + a_t\{r^i_t +$
    $\gamma\pi^i(s_{t+1})\sum_{j=1}^{N} Q^j_t(s_{t+1})\pi^j(s_{t+1})\}$

    where $(\pi^i(s_{t+1}), \pi^j(s_{t+1}))$ are cooperative strategies;
    Let t=t+1;
until rescue complete;
```

**Algorithm 2:** Heuristic action selection

```
function HeuristicActionSelection ($S_t$)
Input  : State $S_t$
Output: best $found\_action$
Choose best actions $A$ based on policy $\pi(S_t)$ and $Q$
$min\_distance = \infty$
for $action_n \in A$ do
    $next\_state_n = perform\_actions(action\_n)$
    $distance$ = HeuristicDistance($next\_state_n$)
    if $distance \leq min\_distance$ then
        $min\_distance = distance$
        $found\_action = action_n$
    end
end
return $found\_action$
```

**Algorithm 3:** Heuristic distance calculation

```
function HeuristicDistance ($S$)
Input  : Current state $S$
Output: heuristic distance at state $S$
- compute distances from agents to victims
- sort distances in ascending
- pick pair matching agent to the shortest victim
- total_distance = sum distance from agents to selected
  victims
return total_distance
```

$\mathrm{A}_1 \times ... \times \mathrm{A}_\mathrm{N} \rightarrow \mathrm{S}$ produces the next state with agents taking the action following the policies $\pi_{\theta_i} : \mathrm{O} \times A_i \rightarrow [0, 1]$. Each agent $i$ receives an observation correlated to the state $o_i : \mathrm{S} \rightarrow \mathrm{O_i}$, and obtains a reward $r_i : \mathrm{S} \times \mathrm{A_i} \rightarrow \mathrm{R}$. Each agent $i$ aims to maximize the shared total expected return $R_i = \sum_{i=1}^{N}\sum_{t=0}^{T}\gamma^t r^t_i$ where $\gamma$ is the discount factor and $T$ is the horizon.

Several reinforcement learning algorithms have been proposed to estimate the value of an action in various contexts. These include the Q-learning, SARSA, and policy gradient algorithm. Among them, the model-free Q-learning algorithm stands out for its simplicity. In Q-learning, the algorithm uses a Q-function to calculate the total reward, defined as $Q : S \times A \rightarrow R$. Q-learning iteratively evaluates the optimal Q-value function using backups:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma max_{a'}Q(s', a') - Q(s, a)] \quad (3)$$

where $\alpha \in [0, 1)$ is the learning rate and the term in the brackets is the temporal-difference (TD) error. Convergence to $Q^{\pi^*}$ is guaranteed in the tabular case provided there is sufficient state/action space exploration.

*2) The Heuristic Function:* The Q-learning requires a number of trials in order to learn and perform consistently, which will increase the total time to generate a rescue plan. In order to address this problem, heuristic-based algorithms have been proposed, e.g. in Robotic Soccer game [27]. For the current problem, we propose a heuristic based Q-learning: ResQ. In our problem, the locations of volunteers and victims will be estimated via tweets' geolocation as described in Section IV-B. We will then incorporate this information as a heuristics function in the learning process. When determining actions for volunteers, besides choosing the optimal Q-value as mentioned earlier, we also prioritize the actions that result in the shortest distance to the victims. The heuristics function is a mapping $H : S \times A \rightarrow R$ where $S$ is the current state, $A$ is the action to be performed, and $R$ is a real number representing the distance of volunteers to the victims. If after performing an action $a$ in $A$, the agent is at row $r_a$ and column $c_a$ of the grid, and its goal is the victim positioned at row $r_v$ and column $c_v$, then the heuristic distance $h$ is calculated as:

$$h = |r_a - r_v| + |c_a - c_v| \quad (4)$$

Our proposed ResQ algorithm for rescue activity is illustrated in Algorithm 1, 2 and 3.

## IV. Experiments and Results

### A. Datasets

Tweets were collected from Aug 23, 2017, until Sept 5, 2017, using the Twitter API, covering the whole course of Hurricane Harvey and its immediate aftermath. The raw data for each day includes about two million tweets, and every tweet has 36 attributes including, among other information, the geographic location, its geographic coordinates, text, and the user profile. The raw data was cleaned by removing tweets that were not tweeted from the United States. Texas state has the largest total number of tweets with scaling the statistics to a range$[1, 1000]$, and the original total number of tweets is 173,315.

### B. Identifying Victims and Volunteers

Victims and volunteers were identified using a series of classifiers. We first designed a classifier to filter Harvey related

TABLE I: Tweet statistics from Aug 23, 2017, to Sept 5, 2017.

| Total tweet | 25,945,502 | Volunteer tweet | 13,953 |
|---|---|---|---|
| Harvey tweet | 173,315 | Victim tweet | 16,535 |

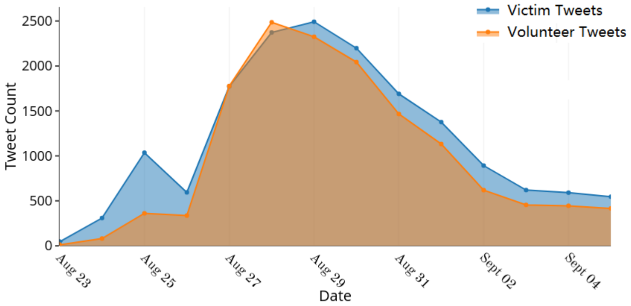
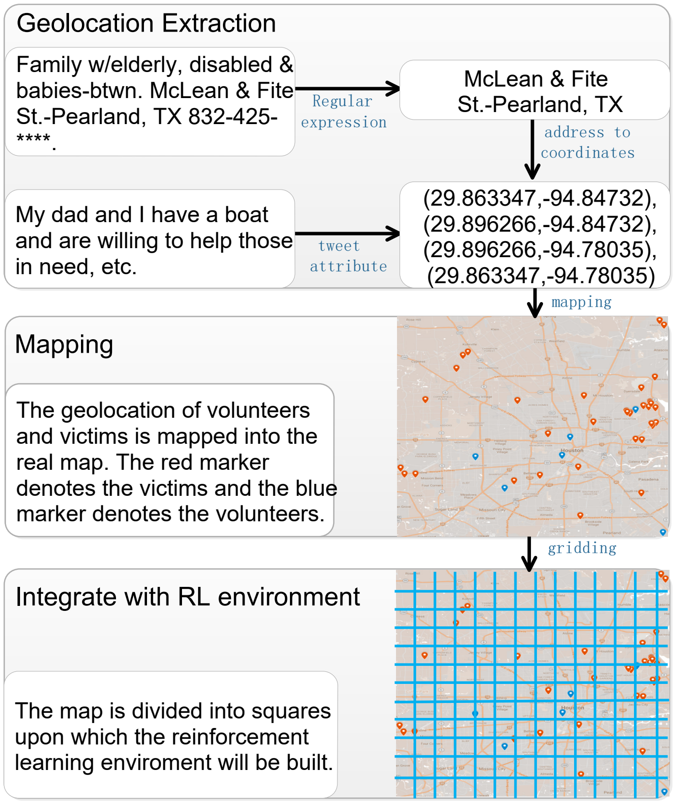


Fig. 3: Time series of victim and volunteer tweet counts.

tweets from all the tweets. In this context, a Harvey tweet refers to a post talking about Hurricane Harvey or related to Hurricane Harvey. In these Harvey tweets, we further developed two classifiers to identify tweets from victims and volunteers. Here, victim tweets are those from victims (or their friends) requesting for help, including retweets. Volunteer tweets are from volunteers who have a boat and are willing to help. All of these classifiers were implemented based on a Support Vector Machine (SVM). In every classifier, $2,000$ tweets were manually labeled, with $80\%$ of the tweets being used for training, and the rest for testing. A five-fold cross-validation method was then applied to ensure the classification results were trustworthy. To obtain a reliable classification result, we compared Logistic Regression, K-Nearest Neighbor (KNN), CART, and SVM. Evaluation metrics such as precision (positive predictive value), recall, F-measure, and accuracy were calculated to evaluate the performance, as shown in Table II.

*1) Victim and Volunteer Time Series:* As shown in Figure 3, we tracked the victims and volunteers tweets time series to monitor the impact of Hurricane Harvey and rescue activities. Initially, when Hurricane Harvey formed into a tropical depression on Aug 23, not much attention was observed from Twitter in the U.S. When Harvey made landfall near the Texas Gulf Coast on Aug 25, there was a burst of victims' tweets. With the increasing of victims requesting help, the number of volunteers also increased sharply and reached a climax on Aug 28. Meanwhile, victims tweets reached a peak on Aug 29. With the leaving of Harvey and system-wide rescuing, both the victims' tweets and volunteers' tweets dropped gradually. Generally, the number of volunteers' tweets is always lower than the victims' tweets.


Geolocation Extraction
Family w/elderly, disabled & babies-btwn. McLean & Fite St.-Pearland, TX 832-425-****.
Regular expression
McLean & Fite St.-Pearland, TX
address to coordinates
My dad and I have a boat and are willing to help those in need, etc.
tweet attribute
(29.863347,-94.84732), (29.896266,-94.84732), (29.896266,-94.78035), (29.863347,-94.78035)
mapping
Mapping
gridding
Integrate with RL environment

The geolocation of volunteers and victims is mapped into the real map. The red marker denotes the victims and the blue marker denotes the volunteers.

The map is divided into squares upon which the reinforcement learning enviroment will be built.

Fig. 4: Reinforcement learning environment transformation.

*2) Geocoding:* To geographically locate victims and volunteers, we designed a simple tool to extract their geolocations. Since geographic coordinates are included in both of tweets from GPS-enabled devices and tweets giving specific addresses, we directly used the address or coordinate to locate the victims or volunteers. Otherwise, we combined alternative sources of information to infer their locations, such as the self-reported location string in the user's profile metadata, or by analyzing the tweet's content. With the help of the World Gazetteer (http://archive.is/srm8P) database, we were able to look up location names and geographic coordinates.

### *C. Experiment Setting*

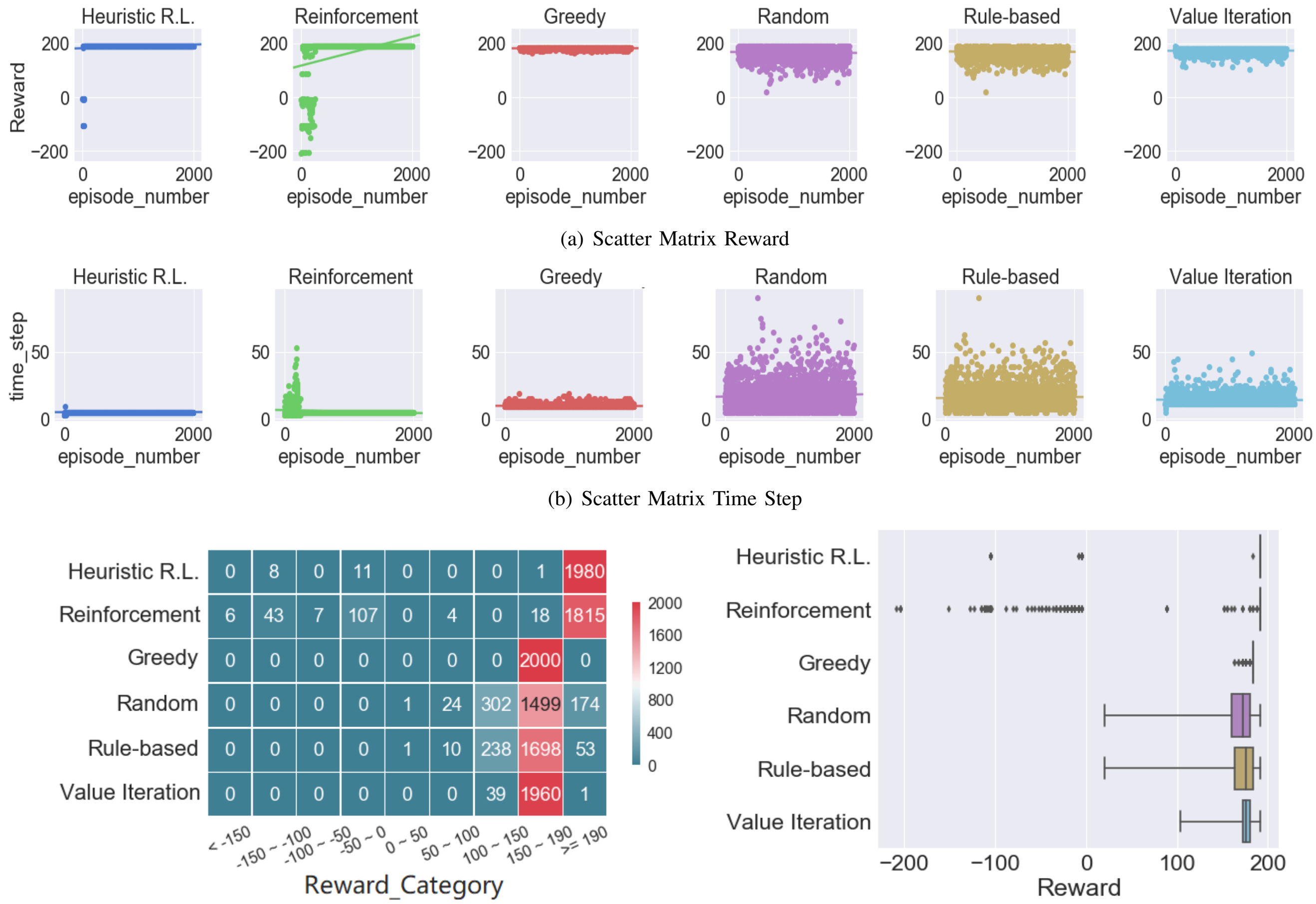
We model the problem of rescue scheduling using a heuristic fully cooperative multi-agent reinforcement learning method. Multi-agent means that we use multiple agents to represent multiple volunteers. The number of agents depends on the number of volunteers identified in the volunteer tweet classification process for each day. Similarly, we assume that the victims are immobile learning targets because victims are trapped. Since volunteers aim to rescue all the victims as soon possible, the goal of all agents is to reach all of their targets with the lowest cost (shortest distance) and maximize the total reward.

In the following sections, we describe how the disaster grid environment is implemented and what actions volunteers can perform in the course of their rescuing activities.

*1) The Grid Environment Identification:* The process of environment building is illustrated in Figure 4. In actual disaster relief operations, the whole city of Houston is the activity space for volunteers, and since a volunteer can go in any direction, the combination of space and direction will

TABLE II: Tweets classification results.

| | Harvey Classification | | | | Victim Classification | | | | Volunteer Classification | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F_ measure** | **Accuracy** | **Precision** | **Recall** | **F_ measure** | **Accuracy** | **Precision** | **Recall** | **F_ measure** | **Accuracy** |
| Log. Regr. | 0.8 | 0.7273 | 0.7619 | 0.8646 | 0.8437 | 0.5510 | 0.6667 | 0.7127 | 0.9583 | 0.6216 | 0.7541 | 0.8170 |
| KNN | 1.0 | 0.2105 | 0.3478 | 0.7580 | 1.0 | 0.8414 | 0.9139 | 0.9172 | 1.0 | 0.6129 | 0.76 | 0.8248 |
| CART | 1.0 | 0.6364 | 0.7778 | 0.8919 | 1.0 | 0.9795 | 0.9896 | 0.9893 | 1.0 | 0.7567 | 0.8645 | 0.8902 |
| SVM | 0.8947 | 0.9444 | 0.9189 | 0.9516 | 0.9146 | 0.9868 | 0.9493 | 0.9490 | 0.9146 | 0.9868 | 0.9493 | 0.9490 |



(a) Scatter Matrix Reward

(b) Scatter Matrix Time Step

(c) Heatmap Reward

(d) Box Plot Reward

Fig. 5: Comparison of the performance of different algorithms.

be infinite. According to our statistics, 95% of the requests for help during the hurricane come from a fixed downtown area. For simplicity, our model is based on a quasi-square area defined by four position coordinates, which are (29.422486,-95.874178), (30.154665,-95.874178), (30.154665,-95.069705) and (29.422486,-95.069705). This square region has a width of 50 miles. For simplicity, the rectangular region is mapped into a 25 by 25 grid, with each grid representing a 4-square-mile area in the real world. By applying this simple mapping to convert the actual map to a virtual grid, we can transform the real world continuous state space to a more manageable discrete state space, and hence significantly reduce the state space complexity.

The position coordinates of the victims and volunteers are extracted every hour following the processes described in the Figure 4. This hourly updating strategy will keep our system updated with the number of available volunteers to be scheduled to rescue the remaining victims, and the number of trapped victims and their positions is also updated. From our observations, For victim tweets that contain the victim's address and phone number, such as *McLean & Fite St.-Pearland, TX 832-425-***** , we can extract the address and converted their position coordinates. For volunteer tweets that do not include the volunteers' address, we can use the geocoding tool as described in section IV-B to extract geographical information from the raw tweet. Once a victim is rescued, we assume the victim is sent to the nearest shelter, and update the numbers and coordinates accordingly.

### D. Disaster Relief Coordination Performance

*1) Baseline Models:* We used the following classical search methods to compare their performances with that of our proposed technique:

*a)* ***Random Walk:*** In this search policy, the agent will randomly walk around the grid and search for any victim they come across along the way. The behavior is random without any other knowledge of the environment.

*b)* ***Greedy Best First Search:*** A greedy best first search offers volunteers a heuristic distance estimation to the victims. Volunteers begin by rescuing the closest victims first and then move on to the further ones sequentially.

*c)* ***Rule-based Search:*** A rule-based search computes action rules by utilizing the probability of taking an action in a grid cell. The action with highest probability are then selected for the next action. This probability is computed from the last average rewards gained in those cells during training episodes controlled by the random walk algorithm. In particular, if $V(t, j)$ is the averaged reward value at time $t$ of the grid cell $g_j$, and the volunteer takes action $a_t$ in order to move to grid cell $g_{j+1}$, the probability of taking action $a_t$ at the grid cell $g_j$ is:

$$p(a_t = [g_j, g_{j+1}]) = \frac{V(t+1, j)}{V(t+1, j) + V(t+1, j+1)} \quad (5)$$

*d)* ***Value Iteration:*** This algorithm works by dynamically updating the value table based on a policy evaluation such as that described by [28]. The allocation policy is computed based on the new value table,

*e)* ***Reinforcement Learning:*** This is traditional reinforcement learning technique where there is no heuristics consideration in action selection. The technique has the same settings such as action, state and reward space compared with our proposed heuristic reinforcement learning.

*2) Evaluation Metrics:* We define an episode as the set of attempts made by all volunteers to successfully rescue all the victims. Hence, our key metrics for measuring the performance of rescue activities are the *average episode time*, *average episode reward*, *average reward rate* and *average rescuing cost*.

*a)* ***Average episode time:*** is the average total time steps required to rescue all the victims in all executed episodes. Each time step is equivalent to one step action (from one cell to another near-by cell) taken by all the available volunteers.

*b)* ***Average episode reward:*** is the average cumulative reward that all volunteers earn in each episode of rescuing.

*c)* ***Average reward rate:*** is the ratio between the average episode reward and average episode time.
$reward\_rate = \frac{\sum_{i=1}^{N} reward_i}{\sum_{i=1}^{N} time_i}$

*d)* ***Average rescuing cost:*** represents the total time step cost to earn one unit of reward. This is the inverse of the reward rate. $rescuing\_cost = \frac{1}{reward\_rate}$

Notice that the model includes the option to have multiple episodes in order to allow us to measure the average performance achieved and the capacity to learn for each rescue policy. Algorithm 1 presents the calculation of the total time steps and total rewards per episode.

TABLE III: Rescue performance comparison. Bold values represent best performance.

| | Time | Reward | Reward Rate | Rescuing Cost |
|---|---|---|---|---|
| Random Walk | 17.4 | 167.2 | 9.6 | 0.104 |
| Greedy B.F.S | 9.6 | 182.7 | 19.03 | 0.053 |
| Rule-based | 15.9 | 170.2 | 10.70 | 0.093 |
| Value Iteration | 14.1 | 173.7 | 12.32 | 0.081 |
| Reinforcement Learning | 5.4 | 172.0 | 31.85 | 0.031 |
| Heuristic R. L. | **5.0** | **189.7** | **37.9** | **0.026** |

*3) Results and Comparisons:* In this work, a heuristic multi-agent reinforcement learning model for disaster relief is trained and evaluated in OpenAI Gym [29]. Unlike the standard reinforcement learning settings used for simulations, our experimental environment setting is based on the real-world geographical positions of tweets. Here, a volunteer is formulated as taking action in an environment and receiving rewards and observation at every time step. The training of the agent stops once the policies of volunteers converge. The main purpose is to minimize the amount of time needed to rescue all the victims in the target environment.

For these experiments, we transform the geographical distribution of tweets into a grid and set up a centralized communication environment, which consists of N volunteers and M victims in a two-dimensional grid with discrete space and discrete time. The process of extracting geographical information from volunteers and victims is illustrated in Figure 4. Volunteers may take actions in the environment and communicate with the remote central server. They will be assigned a penalty if they go off the grid and a reward if they reach the victims they are to rescue.

We compared the experimental performance of the proposed ResQ algorithm with Random walk, Greedy best first search, Rule-based search, Value iteration, and a traditional Reinforcement Learning method. Figure 5 presents the process of each algorithm's performance within 2000 episode (path from initial to a terminal state). In Figure 5(a) and Figure 5(b), we compare the total rewards and total time steps per episode with each strategy. The ResQ quickly converges to stable states after the first 24 episodes of training. Once ResQ converged, it constantly outperforms all other approaches. As a comparison, the reinforcement learning technique also performs well after convergence. However, it requires a long time for convergence (208 episodes in current experiment) and the average reward over the entire time period is lower compared to the ResQ. The greedy B.F.S strategy performs consistently over the time, shown as points around constant lines. This is not surprising because with this strategy the agents always choose to reach the closest victims first, which is independent of other factors in the rescuing environment. Overall, the reward of greedy B.F.S strategy is less than the ResQ, while its time steps outperform the ResQ during the latter's training phase. The Random walk approach leads to the lowest overall reward as well as the highest completion time per episode, and the performance has a large variation across different episodes. Ruled based and Value iteration are even worse compared to our proposed ResQ technique. Figure5(c) and Figure5(d) respectively show the heatmap of the reward distribution and its corresponding box plot. We clearly see that, during the total 2000 testing episodes, the ResQ has the most of its rewards above 190, while other methods have significantly less number of rewards in this category.

Table III gives a summary of each algorithm's total time, total reward, reward rate, and rescuing cost. The result clearly shows that the ResQ has the best overall reward score, the shortest completion time, the highest reward rate, and the lowest rescuing cost rate. In particular, the Greedy B.F.S. and the Reinforced Learning method respectively have the reward and time performance close to the proposed method. Nonetheless, the proposed Heuristic reinforcement learning evidently outperforms these methods when the two metrics are considered simultaneously.

## V. Discussion

This paper presents a novel algorithm designed to develop a better response to victims' requests for assistance during disasters, along with a case study using Twitter data collected during Hurricane Harvey in 2017. This work is one of the first attempts to formulate the large-scale disaster rescue problem as a feasible heuristic multi-agent reinforcement learning problem using massive social media data. With the proposed method, we can train classifiers to extract the victim and volunteer information from tweets and transform the data for use in a reinforcement learning environment. Our key contribution is the design of a heuristic multi-agent reinforcement learning scheduling policy that simultaneously schedules multiple volunteers to rescue disaster victims quickly and effectively. The experimental results showed that the heuristic multi-agent reinforcement learning algorithm could respond to dynamic requests and achieve an optimal performance over space and time. Also, the results showed that this approach could match volunteers and victims for faster disaster relief and better use of limited public resources. The proposed framework for disaster exploration and relief recommendation is significant in that it provides a new disaster relief channel that can serve as a backup plan when traditional helplines are overloaded.

## VI. Acknowledgment

This work was supported by the U.S. National Science Foundation (NSF) under Grant CNS-1737634.

## References

[1] X. Zhu, "Semi-supervised learning literature survey," *world*, 2005.

[2] Z. Yang, L. H. Nguyen, J. Stuve, G. Cao, and F. Jin, "Harvey flooding rescue in social media," in *2017 IEEE Big Data (Big Data)*. IEEE, 2017, pp. 2177–2185.

[3] Z. Gong and et al., "Press: Predictive elastic resource scaling for cloud systems," in *2010 CNSM*, 2010, pp. 9–16.

[4] W. Song and et al, "Adaptive resource provisioning for the cloud using online bin packing," *IEEE Transactions on Computers*, 2014.

[5] W. Iqbal, M. N. Dailey, D. Carrera, and P. Janecek, "Adaptive resource provisioning for read intensive multi-tier applications in the cloud," *Future Generation Computer Systems*, 2011.

[6] Y. Jiang and et al., "Asap: A self-adaptive prediction system for instant cloud resource demand provisioning," in *2011 IEEE 11th ICDM*, 2011.

[7] H. Lu and et al., "Motor anomaly detection for unmanned aerial vehicles using reinforcement learning," *IEEE internet of things journal*, pp. 2315–2322, 2018.

[8] A. Nagabandi and et al., "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *2018 IEEE ICRA*. IEEE, 2018, pp. 7559–7566.

[9] A. Nair and et al., "Overcoming exploration in reinforcement learning with demonstrations," in *2018 IEEE ICRA*. IEEE, 2018, pp. 6292–6299.

[10] T. Nazer and et al., "Intelligent disaster response via social media analysis a survey," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 46–59, 2017.

[11] M. Jamali, A. Nejat, S. Ghosh, F. Jin, and G. Cao, "Social media data and post-disaster recovery," *International Journal of Information Management*, vol. 44, pp. 25–37, 2019.

[12] H. Gao, G. Barbier, and R. Goolsby, "Harnessing the crowdsourcing power of social media for disaster relief," *IEEE Intelligent Systems*, vol. 26, no. 3, pp. 10–14, 2011.

[13] H. Du, L. Nguyen, Z. Yang, H. Abu-Gellban, X. Zhou, W. Xing, G. Cao, and F. Jin, "Twitter vs news: Concern analysis of the 2018 california wildfire event," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2. IEEE, 2019, pp. 207–212.

[14] L. Nguyen, Z. Yang, J. Li, Z. Pan, G. Cao, and F. Jin, "Forecasting people's needs in hurricane events from social network," *IEEE Transactions on Big Data*, 2019.

[15] L. H. Nguyen, S. Jiang, H. Abu-gellban, H. Du, and F. Jin, "Nipred: Need predictor for hurricane disaster relief," in *Proc. of the 16th SSTD*, 2019, pp. 190–193.

[16] L. H. Nguyen, A. Salopek, L. Zhao, and F. Jin, "A natural language normalization approach to enhance social media text reasoning," in *2017 IEEE Big Data*. IEEE, 2017, pp. 2019–2026.

[17] M. Zook and et al., "Volunteered geographic information and crowdsourcing disaster relief: A case study of the haitian earthquake," *World Medical & Health Policy*, pp. 7–33, 2010.

[18] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, "Tweedr: Mining twitter to inform disaster response," in *Proc. ISCRAM'14*. ISCRAM Association, 2014.

[19] J. Hu and M. P. Wellman, "Nash q-learning for general-sum stochastic games," *Journal of Machine Learning Research*, vol. 4, pp. 1039–1069, 2003.

[20] R. Powers and Y. Shoham, "New criteria and a new algorithm for learning in multi-agent systems," in *Advances in Neural Information Processing Systems 17*, 2005.

[21] M. Bowling, "Convergence and no-regret in multiagent learning," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2005, pp. 209–216.

[22] M. Bowling and M. Veloso, "Rational and convergent learning in stochastic games," in *Proc. IJCAI'01*, ser. IJCAI'01, 2001, pp. 1021–1026.

[23] M. L. Littman, "Value-function reinforcement learning in markov games," *Cognitive Systems Research*, vol. 2, no. 1, pp. 55 – 66, 2001.

[24] S. Liang, Z. Yang, F. Jin, and Y. Chen, "Data centers job scheduling with deep reinforcement learning," in *PAKDD*. Springer, 2020, pp. 906–917.

[25] R. Bellman, *Dynamic Programming*, 1st ed. Princeton University Press, 1957.

[26] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. John Wiley & Sons, Inc., 1994.

[27] R. A. Bianchi, C. H. Ribeiro, and A. H. R. Costa, "Heuristic selection of actions in multiagent reinforcement learning." in *IJCAI*, 2007, pp. 690–695.

[28] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1.

[29] G. Brockman and et al., "Openai gym," 2016.