

Clustering of Social Media Messages for Humanitarian Aid Response during Crisis

Swati Padhee^{1*}, Tanay Kumar Saha², Joel Tetreault² and Alejandro Jaimes²

¹Wright State University, Dayton, OH

²Dataminr Inc., New York, NY

padhee.2@wright.edu, {tsaha, jtetreault, ajaimes}@dataminr.com

Abstract

Social media has quickly grown into an essential tool for people to communicate and express their needs during crisis events. Prior work in analyzing social media data for crisis management has focused primarily on automatically identifying actionable (or, informative) crisis-related messages. In this work, we show that recent advances in Deep Learning and Natural Language Processing outperform prior approaches for the task of classifying informativeness and encourage the field to adopt them for their research or even deployment. We also extend these methods to two sub-tasks of informativeness and find that the Deep Learning methods are effective here as well.

1 Introduction

During large-scale disasters, humanitarian organizations seek timely and reliable information to understand the overall impact of the crisis in order to respond effectively. Social media platforms, such as Twitter, have emerged as an essential source of timely, on-the-ground information for first responders [Vieweg *et al.*, 2014; Castillo, 2016]. Public Information Officers utilize online social media to gather actionable crisis-related information and provide it to the respective humanitarian response organizations [Castillo, 2016].

Manually monitoring and classifying millions of incoming social media messages during a crisis is not possible in a short time frame as most organizations lack the resources and workforce. Thus, it is necessary to leverage *automatic* methods to identify informative messages to assist humanitarian organizations and crisis managers. An example of an informative message on Twitter during Hurricane Dorian is: “All Tidewater Dental Locations are collecting Donations for the Victims of Hurricane Dorian! Items needed: (PLEASE SHARE) Non-perishable foods, Bug Repellent, Blankets, Clean clothing, Socks, Wipes, Toiletries”. To date, most works in this area have focused on leveraging classical machine learning methods [Imran *et al.*, 2014;

Caragea *et al.*, 2016; Nguyen *et al.*, 2016] for detecting informative messages at a reasonable accuracy.

Our paper makes two contributions. First, we show that recent Deep Learning and Natural Language Processing methods, specifically state-of-the-art pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers) [Devlin *et al.*, 2018] and RoBERTa [Liu *et al.*, 2019], yield substantial improvements in this task and encourage the field to adopt them. We also analyze how robust these methods are to new crisis events.

Second, prior work has tended to focus on the binary classification task of whether a message is informative or not. However, more refined classifications can be of further assistance. We propose two new classification tasks motivated by the need of humanitarian organizations to cluster actionable crisis-related messages into their *intent* (“Need” and/or “Supply”), and *humanitarian aid type* (“Food”, “Shelter”, “Health”, and “WASH”), where WASH stands for (Water, Sanitation, and Hygiene). The latter task is based on the UN Humanitarian Reform process, which outlines eleven primary needs or clusters that humanitarian organizations should track during an emergency. Figure 1 presents an overview of our proposed architecture for the UN cluster motivated humanitarian aid response management.

In short, we show that recent neural approaches offer substantial gains across three humanitarian aid classification tasks, of which two are new. The analyses above can be applied to any social media platform, but for this research study, we use messages posted on Twitter during Hurricane Dorian, a Category 5 hurricane that impacted North America in 2019. In Section 3, we detail the process for constructing a dataset for each of the three tasks. In Section 4, we describe the models, experimental setup, and results. Finally, in Section 4.3, we discuss how well the top model performs on unseen crisis events and conclude with future directions in Section 5.

We provide our keywords¹ and annotation guidelines to the community to make our study reproducible.

2 Related Work

The task of detecting informative messages from Twitter, or other social media, is typically treated as a three-step process. First, simple filters are employed to extract tweets rel-

*Research work was done while Swati was interning at Dataminr Inc.

¹<https://github.com/swatipadhee/Crisis-Aid-Terms.git>

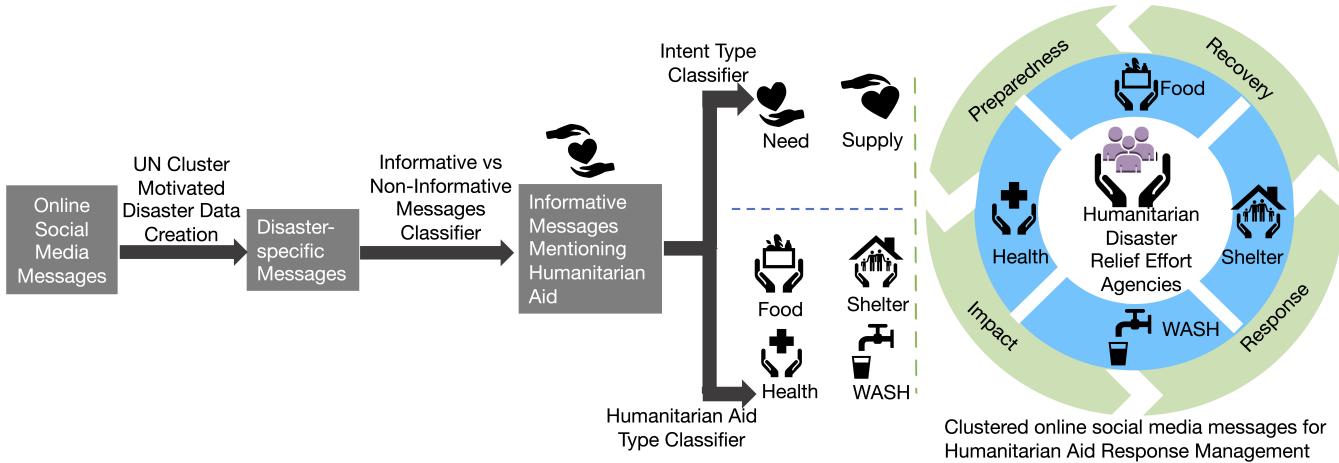


Figure 1: Overview of the UN Cluster motivated Humanitarian Aid Response Management

event to the event. Second, machine learning methods are developed to extract informative tweets from the first step automatically. A third step may follow in which the informative tweets are labeled with more specific distinctions of informativeness. We detail the related work for these steps below.

Extracting disaster-specific social media messages usually relies on two types of keyword matching: (a) *lexicons*, or keywords related to the disaster, to extract tweets mentioning disaster keywords [Purohit *et al.*, 2014; Hodas *et al.*, 2015; Imran *et al.*, 2013], and (b) *location terms* to extract all tweets associated with the areas impacted by the disaster [Mahmud *et al.*, 2012; Waqas and Imran]. The lexicon-based approach introduces noise (low precision), while the location-based approach has shallow coverage (low recall) and might fail to capture a considerable fraction of relevant tweets. We use both lexicon and location-specific terms in collecting tweets from the Hurricane Dorian that hit in 2019.

Prior works have filtered actionable, informative tweets from large datasets during disasters [Nguyen *et al.*, 2015; Caragea *et al.*, 2016; Zhang and Vucetic, 2016] employing classical machine learning as well as deep learning-based techniques for the classification task [Madichetty and Sridevi, 2019; Imran *et al.*, 2014; Caragea *et al.*, 2016; Zhang and Vucetic, 2016; Caragea *et al.*, 2016; Neppalli *et al.*, 2018; Nguyen *et al.*, 2016]. More recently, deep learning methods have been successful for a host of NLP tasks such as summarization [Wu and Hu, 2018], machine translation [Edunov *et al.*, 2018], Named entity recognition [Baevski *et al.*, 2019] etc. The amount of work leveraging deep learning methods for emergency response tasks is limited. Of note are Jain *et al.* [2019], which experiments with embeddings such as BERT Devlin *et al.* [2018], ELMo Peters *et al.* [2018], GloVe Pennington *et al.* [2014] and word2vec Mikolov *et al.* [2013], and Alam *et al.* [2020] which experiments with CNNs and BERT. Informativeness classification varies based on the dataset, but it is possible to achieve performance as high as 0.87 F-score Alam *et al.* [2020]. In our work, we experiment with a range of machine learning and deep learning methods, including a recent one, RoBERTa [Liu *et al.*, 2019].

Finally, there is some work into further classifying informative messages from the prior stage. Examples of categories include caution or advice, information source, people, casualties, damage, and donations [Imran *et al.*, 2013; Neppalli *et al.*, 2018; Alam *et al.*, 2020; Madichetty and Sridevi, 2019; Maas *et al.*, 2019]². The Multilingual Disaster Response Dataset³ covers 36 humanitarian categories, including Food, Shelter, Water, Clothing, Medical Help, and Medical Products. The humanitarian categorizations of this dataset do not directly align with the need of UN cluster-specific humanitarian organizations, and also, the annotations are not granular enough to differentiate between “Need” and “Supply”. Similar to informativeness classification, classical machine learning, and deep learning algorithms [Alam *et al.*, 2020; Jain *et al.*, 2019] are mostly used for humanitarian task type classification. We use similar methods for our tasks and show that a recent state-of-the-art language model, RoBERTa performs the best in informativeness classification as well as the newly proposed tasks.

3 Data Creation

3.1 Unsupervised Data Extraction for Hurricane Dorian

We utilize humanitarian help-type keywords and disaster-specific location terms to extract tweets posted during Hurricane Dorian from Aug 24, 2019, to Sep 23, 2019. We design custom queries by combining the generic disaster-specific keywords used in previous work [Alam *et al.*, 2018a,b; Olteanu *et al.*, 2014] along with: (a) UN cluster-based lexical keywords released in [Temnikova *et al.*, 2015], (b) humanitarian aid type-specific keywords [Niles *et al.*, 2019], and (c) generic disaster location terms (e.g., “Bahamas”). We utilize a publicly available Python library (GetOldTweets⁴) with

²For a more detailed summary of the different categories and datasets, we refer the reader to [Alam *et al.*, 2020], Section 8.2

³<https://appen.com/datasets/combined-disaster-response-data/>

⁴<https://pypi.org/project/GetOldTweets3/>

Models	Informativeness Classification Total: 1,208 Train: 966; Test: 242		Intent Type Classification Total: 1,010 Train: 808; Test: 202		Aid Type Classification Total: 1,010 Train: 808; Test: 202			
			Need F1 (%)	Supply F1 (%)	Food F1 (%)	Shelter F1 (%)	Health F1 (%)	WASH F1 (%)
	Acc (%)	F1 (%)						
MNB	83.67	83.32	57.69	62.67	77.74	60.25	74.14	65.42
LR	77.67	77.78	62.87	67.38	88.59	63.32	74.15	64.39
BERT	85.67	87.81	67.57	65.39	74.74	63.19	74.83	80.77
RoBERTa	87.67	89.20	71.98	75.98	89.53	69.02	78.89	85.81

Table 1: Categorization Results of Social Media Messages for Humanitarian Aid Response

those custom queries to extract relevant tweets during Hurricane Dorian. This process results in 37,768 unique tweets that serve as our dataset for results and discussion. We remove all user information from the tweets (i.e., handles).

3.2 Labeled Dataset Creation for Three Humanitarian Tasks

Next, we use the tweets from above to build three labeled datasets for three respective supervised tasks: Informativeness, Intent Type, and Aid Type. For each dataset, we use the Amazon Mechanical Turk (AMT) platform⁵ to generate ground-truth labels, which we compare system predictions to. Annotation instructions for each task can be found in Table 2.

Task 1: Informativeness Classification:

We sample a set of 1,208 tweets from the collected 37,768 unique tweets (we remove tweets with cosine similarity higher than 0.85) uniformly at random and employ paid expert workers from AMT to generate ground-truth labels of informativeness (whether the tweet is informative or not). Among 1,208 tweets, 482 (39.91%) were labeled as Informative.

Tasks 2 & 3: Intent Type and Aid Type Classification:

Based on the labeled dataset for Task 1, we develop a binary classification model (please see Section 4.2 for detailed results) and use the RoBERTa model to predict labels on the remaining tweets. The model predicted 14,073 (37.26%) out of 37,768 tweets as informative. For Tasks 2 and 3, we sample 1,010 tweets uniformly at random from those 14k informative tweets. For each task and tweet, we got labels from 5 AMT “master level” annotators. We decide the final labels for each tweet based on the agreement of three or more annotators. To measure the quality of the annotations, the authors manually annotated 300 tweets and observed a substantial agreement with the majority label produced by the AMT annotators (Cohen’s Kappa score [Cohen, 1960] of 0.71). We focus on the following humanitarian aid types: food, shelter (temporary or permanent home, basic living needs like clothes or electricity, etc.), water, sanitation, hygiene, and health support. For Task 2, AMT workers labeled 781 (77.33%) tweets as “Need” and 470 (46.54%) as “Supply”. As expected, “Need” tweets are more prevalent than “Supply” tweets. For Task 3, AMT workers labeled 470

(46.54%) as “Food”, 291 (28.81%) as “Shelter”, 164 (16.24 %) as “Health” and 276 (27.33 %) as “WASH”.

4 Experiments

With the three datasets in place, we can benchmark different modeling approaches head to head. For each dataset, we remove URLs, image links, numbers, hashtags, mentions, non-ASCII characters from tweets, and contract multiple spaces into a single space. We use Micro-F1 for binary classification task (Task 1), and Macro-F1 for multi-label multi-class tasks (Task 2 and Task 3). We also report accuracy for the informativeness classification task. We split each dataset into train (80%) and test (20%) partitions. We run our experiments five times and report the average value for each metric.

4.1 Models Compared

We present our tasks’ performance on two baseline traditional machine learning algorithms using TF-IDF embeddings and a linear softmax classifier using two contextual language model embeddings.

Multinomial Naïve Bayes (MNB) is a learning technique built upon the theory that the features representing the data points are conditionally independent of each other for a given class. We use TF-IDF based embeddings of documents as features for this model.

LR is designed using a linear classification function that predicts the probability of a data belonging to a particular class. We use TF-IDF based embeddings of documents as features for this model. We use the Scikit-learn pipeline to generate TF-IDF vector representations⁶.

BERT [Devlin *et al.*, 2018] is a document representation learning model which looks into both left and right context of a word to learn the representation. We choose the pre-trained BERT-base model and add a task-specific fine-tuning layer on top of the BERT architecture for the classification tasks.

RoBERTa [Liu *et al.*, 2019] is a language model similar to BERT, but, trained by modifying the design strategies in BERT to achieve better performance in downstream tasks. We add a task-specific fine-tuning layer on top of the RoBERTa architecture for the classification task.

⁵<https://www.mturk.com>

⁶<https://scikitlearn.org/>

Task	Instructions
Task 1: Informativeness	Given a tweet, select "Yes" if the tweet is talking about either people requesting humanitarian help during a hurricane or that help is on the way. Humanitarian help includes food, shelter, water, hygiene, mental, or physical health support. Select "No" if the tweet is not talking about any of the humanitarian help types.
Task 2: Intent Type	Select "Need" if a tweet contains a mention of the need for humanitarian help, regardless of who mentions it and for whom. Select "Supply" if a tweet contains a mention of the supply of humanitarian help. Select "Both" if, in a tweet, there is mention of both need and supply. If the tweet is NOT about either need or supply, please select "None of the above."
Task 3: Aid Type:	If there is only one help type, pick one. If there are multiple help types, pick all of them that are relevant. If a tweet is NOT about any of the help types, please select "None of the above." The choices are: "Food," "Shelter," "Health," "Water, Sanitation, and Hygiene (WASH)," or "None of the above."

Table 2: Annotation Instructions for Mechanical Turk

Tweet texts	Human Label	RoBERTa
A Tamil-English translator needed. #FloodSL #SriLanka SFHS English Department has you covered with your back to school, post Harvey, supply needs! Stop by D103! Meet Irma!! The go-to LulaRoe top. It is loose, knit high-lo tunic with fitted sleeves!#lularoe #lularoeirma	Non-Informative Informative Informative	Non-Informative Informative Non-Informative

Table 3: Informativeness Classification on Other Crisis Events Datasets

4.2 Results

Task 1 Results: Table 1 (First part) reports the results for our first classification task (informativeness vs non-informativeness). In this task, RoBERTa outperforms all other models. The model achieves an accuracy of 87.67%, and an F1 score of 89.20%. The closest competitor of RoBERTa is BERT, which is also based on contextual embeddings. These results indicate that contextual embeddings have much better discriminative power than the traditional TF-IDF based embeddings.

Task 2 Results: Table 1 (Second Part) reports the results for the second humanitarian task of classifying a tweet as either Need or Supply. Again, RoBERTa performs best with a Micro-F1 measure of 71.98% for predicting "Need" and that of 75.98% for predicting "Supply." Similar to Task 1, BERT is the second best model.

Task 3 Results: Table 1 (Third Part) shows the results for the classification of UN cluster-based humanitarian categories. RoBERTa achieves the highest accuracy and Micro-F1 across all four categories by a wide margin.

We then use RoBERTa to predict labels on all of the 14,073 informative tweets filtered using the classifier trained in Task 1. Although our proposed Tasks 2 and 3 are fine intent-driven categorization of humanitarian aid types, we observe results with respect to BERT in the same spirit of Jain *et al.* [2019] where they reported BERT performing poorer than Word2Vec or Glove embeddings in information-type classification. Out of the 14,073 informative tweets, RoBERTa predicted 8,370 (59.48%) tweets seeking for help and 9,296 (66.1%) tweets supplying help. As a given tweet can mention both need and supply messages, the model predicted both the labels for 3,593 tweets. On the same 14,073 informative tweets, RoBERTa model predicted Food as a label for 7,940 (56.42%) tweets, Shelter as a label for 3,543 (25.2%), 2359 (16.76%) as Health, and 4,362 (30.1%) as WASH. The percentage of tweets predicted correctly reflects the significance of tweets to be channeled to the respective humanitarian organizations.

4.3 Discussion

We analyze how a model trained on one crisis event could generalize to unseen crisis events. In an ideal case, such a model would perform well across different crisis events, and no additional data collection and retraining is necessary to deploy. In our case, we investigate how well an Informative classifier trained on Hurricane Dorian can perform on other crisis datasets for the Informativeness task. We use a publicly available dataset of tweets collected during Hurricane Harvey, Hurricane Maria, and Sri Lanka floods [Alam *et al.*, 2018a]. RoBERTa achieved an accuracy of 82.12% in "Sri Lanka Floods", but performed poorly on "Hurricane Harvey" and "Hurricane Maria" (achieved 54.30% and 59.85% accuracy, respectively). These low results indicate that we need a better strategy for domain adaptation and also require adjustment of labels. In Table 3, we list a set of predictions using our best informativeness classifier (RoBERTa) on other crisis datasets [Alam *et al.*, 2018a]. Interestingly, RoBERTa classified the third tweet (third row) in Table 3 as "Non-Informative" contrary to the human-provided label. Nonetheless, the predicted label is correct based on our task definition (Section 3.2) because the tweet is not talking about any humanitarian help and thus "Non-Informative".

5 Conclusion & Future Work

In this work, we show that recent neural approaches offer substantial gains across three humanitarian aid classification tasks. We introduce two additional levels of abstraction (UN Cluster motivated clustering) on top of informativeness classification and show with high accuracy that these tasks can be automated, which can be beneficial to the humanitarian organizations in assessing, prioritizing, and mobilizing the needs of the affected community. Our results show that specifically state-of-the-art language models such as RoBERTa [Liu *et al.*, 2019], yield substantial improvements in these tasks, and we encourage the field to adopt them. In the future, we plan for a qualitative evaluation of our tasks by showing the clustered tweets to the personnel from the respective UN clusters.

References

- Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismm: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*, June 2018.
- Firoj Alam, Ferda Ofli, Muhammad Imran, and Michael Aupetit. A twitter tale of three hurricanes: Harvey, irma, and maria. *Proc. of ISCRAM, Rochester, USA*, 2018.
- Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. Standardizing and benchmarking crisis-related social media datasets for humanitarian information processing. *arXiv preprint arXiv:2004.06774*, 2020.
- Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*, 2019.
- Cornelia Caragea, Adrian Silvescu, and Andrea H Tapia. Identifying informative messages in disaster events using convolutional neural networks. In *International Conference on Information Systems for Crisis Response and Management*, pages 137–147, 2016.
- Carlos Castillo. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press, 2016.
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.
- Nathan O Hodas, Greg Ver Steeg, Joshua Harrison, Satish Chikkagoudar, Eric Bell, and Courtney D Corley. Disentangling the lexicons of disaster response in twitter. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1201–1204, 2015.
- Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1021–1024. ACM, 2013.
- Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 159–162. ACM, 2014.
- Pallavi Jain, Robert Ross, and Bianca Schoen-Phelan. Estimating distributed representation performance in disaster-related social media classification. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 723–727, 2019.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Paige Maas, Shankar Iyer, Andreas Gros, Wonhee Park, Laura McGorman, Chaya Nayak, and P Alex Dow. Facebook disaster maps: Aggregate insights for crisis response & recovery. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3173–3173. ACM, 2019.
- Sreenivasulu Madichetty and M Sridevi. Detecting informative tweets during disaster using deep neural networks. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, pages 709–713. IEEE, 2019.
- Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In *ISCRAM*, 2018.
- Minh-Tien Nguyen, Asanobu Kitamoto, and Tri-Thanh Nguyen. Tsum4act: A framework for retrieving and summarizing actionable tweets during a disaster for reaction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 64–75. Springer, 2015.
- Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Rapid classification of crisis-related data on social networks using convolutional neural networks. *arXiv preprint arXiv:1608.03902*, 2016.
- Meredith T Niles, Benjamin F Emery, Andrew J Reagan, Peter Sheridan Dodds, and Christopher M Danforth. Social media usage patterns during natural hazards. *PloS one*, 14(2):e0210484, 2019.
- Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*, 19(1), 2014.

Irina P Temnikova, Carlos Castillo, and Sarah Vieweg. Emterms 1.0: A terminological resource for crisis tweets. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*, 2015.

Sarah Vieweg, Carlos Castillo, and Muhammad Imran. Integrating social media communications into the rapid assessment of sudden onset disasters. In *International Conference on Social Informatics*, pages 444–461. Springer, 2014.

Humaira Waqas and Muhammad Imran. Campfiremissing: An analysis of tweets about missing and found people from California wildfires.

Yuxiang Wu and Baotian Hu. Learning to extract coherent summary via deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Shanshan Zhang and Slobodan Vucetic. Semi-supervised discovery of informative tweets during the emerging disasters. *arXiv preprint arXiv:1610.03750*, 2016.