"Please Donate for the Affected": Supporting Emergency Managers in Finding Volunteers and Donations in Twitter Across Disasters

Pooneh Mousavi
University of Maryland, College Park
poonehm@umd.edu

Cody Buntain
University of Maryland, College Park
cbuntain@umd.edu

ABSTRACT

Despite the outpouring of social support posted to social media channels in the aftermath of disaster, finding and managing content that can translate into community relief, donations, volunteering, or other recovery support is difficult due to the lack of sufficient annotated data around volunteerism. This paper outlines three experiments to alleviate these difficulties. First, we estimate to what degree volunteerism content from one crisis is transferable to another by evaluating the consistency of language in volunteer- and donation-related social media content across 78 disasters. Second it introduces methods for providing computational support in this emergency support function and developing semi-automated models for classifying volunteer- and donation-related social media content in new disaster events. Results show volunteer- and donation-related social media content is sufficiently similar across disasters and disaster-types to warrant transferring models across disasters, and we evaluate simple resampling techniques for tuning these models. We then introduce and evaluate a weak-supervision approach to integrate domain knowledge from emergency response officers with machine learning models to improve classification accuracy and accelerate this emergency support in new events. This method helps to overcome the scarcity in data that we observe related to volunteer- and donation-related social media content.

Keywords

social media, crisis informatics, volunteers, donations, emergency support functions

INTRODUCTION

Following major disasters, social media channels are consistently inundated with messages of sentiment and support for those affected (C. L. Buntain and Lim 2018; Olteanu, Vieweg, et al. 2015). In crisis informatics research, however, such social support is generally dismissed as low priority, with systems instead focusing on actionable content e.g., (Purohit et al. 2018; Mccreadie et al. 2019; McCreadie et al. 2020) that increases situational awareness during and in the immediate aftermath of a crisis event. Despite this research priority, post-crisis recovery – e.g., connecting potential volunteers to volunteer opportunities, identifying relief efforts, or disseminating information about available shelters and aid – can benefit from better information sharing and retrieval methods in social media (Mccreadie et al. 2019; McCreadie et al. 2020; Glasgow et al. 2016). In fact, this discovery and coordination of volunteers is a critical aspect of disaster response, as highlighted in Emergency Support Function (ESF) #6 on mass care, assistance, and human services in the National Response Framework released by the US Federal Emergency Management Agency (FEMA) (FEMA 2016). Consequently, limited research exists to aid disaster-management personnel and other stakeholders in using social media for coordinating volunteers and donations during the post-crisis recovery phase. This paper takes a step toward addressing this gap by 1) evaluating how well common assumptions in crisis informatics – specifically, consistency in language across disasters – hold in the context of post-crisis recovery and by 2) introducing a method for integrating domain expertise into language classification models that identify volunteer-related social media messaging during crises.

To examine common crisis-informatics assumptions, we first highlight that much of the research around crisis-oriented information systems assumes that discussions across multiple crises are sufficiently similar that insights

from one crisis are transferable to another (Olteanu, Castillo, et al. 2014; C. L. Buntain and Lim 2018; McCreadie et al. 2020; S. Ghosh, K. Ghosh, Chakraborty, et al. 2017). While prior work suggests this assumption holds for expressions of sentiment and other types of information (C. L. Buntain and Lim 2018; Olteanu, Vieweg, et al. 2015), FEMA's ESF6 documentation highlights the importance of hyper-local information in coordinating mass-care capabilities, and such information may transfer poorly across crises. This paper therefore empirically evaluates this consistency specific to the context of volunteer- and donation-related social media content – to which we refer as volunteerism discourse – by comparing language models across 78 disasters made available as part of the of the Incident Streams track at the annual Text Retreival Conference (TREC-IS) (Mccreadie et al. 2019; McCreadie et al. 2020). Using a combination of text analysis and text classification models coupled with algorithmic approaches to resampling and improving robustness of our statistical estimates, our first study demonstrates sufficient similarities in volunteer and donation content exist across 78 crises of 13 different types.

Building on these results and ESF6's guidance on the role of nongovernmental organizations (NGOs) in post-crisis recovery, we explore methods for integrating domain knowledge about NGOs and their online presence to improve classification of volunteer and donation content. In particular, we develop and evaluate several computer-supported strategies and semi-automated models for classifying this social media content in new disaster events by applying label-spreading and semi-supervision approaches from machine learning to samples of disaster-related content and content collected from well-established NGO Twitter accounts. Results show that this hybrid, semi-supervised approach for combining machine learning and domain knowledge significantly improves performance in identifying volunteer and donation content in new crisis events.

Research Questions We have three principal research questions about applicability of recovery- and volunteer-related social media messages across disasters of varying types:

- **RQ1**: How similar is volunteerism and recovery language across different events and event-types?
- **RQ2**: Are models for classifying volunteer-related information transferable from one crisis to another especially if they come from a different context? And what are the best sampling strategies to train a generic model that could be applied for identifying volunteerism content across different events?
- **RQ3**: How might we improve models for identifying recovery and volunteerism content by enriching our labeled dataset with domain experts' insights without the need for time-consuming and labor-intensive annotation?

Contributions This work's primary contributions should be of value to two groups: i) crisis-informatics researchers who want to integrate social media into computer-supported processes for supporting post-crisis recovery, and ii) NGOs and donation platforms who may want to identify potential volunteers or sources of donation-and-recovery activity in social media. In particular, we see this work as presenting the following contributions for these groups:

- An analysis of consistencies in language about volunteerism and donations across disasters and event types,
- Experiments with and identification of good strategies for re-weighting crisis events of similar types when learning the volunteer/donation information, and
- Best practices for weak-supervision strategies that incorporate crisis-response domain knowledge into the learning task.

RELATED WORK

This paper touches on three main areas of related work: First, several research efforts have explored the role of social media in post-disaster community recovery, and we outline how this work supports and builds on such studies. Second, crisis informatics has provided a significant body of computational methods for identifying relevant and informative messages in the lead-up and during crises, and we discuss how we leverage these efforts for data collection and evaluation. Lastly, our efforts to provide computational support for identifying volunteerism content in social media overlaps with research from the text mining, which we briefly outline.

Social Media's Role in Post-Disaster Community Recovery

Prior studies have demonstrated how online spaces support volunteer coordination, especially in the aftermath of disaster. Early studies of online crisis communication, such as Shklovski et al. 2008, have shown pervasive information and communication technology (ICT) facilitate reconnection among communities who have impacted by disaster and accelerate community relief, donations, volunteering, and other recovery support. Numerous related studies (e.g., Starbird and Palen 2013, St. Denis et al. 2012, Cobb et al. 2014, White et al. 2014) all discuss the role of virtual, online volunteers organization such as Humanity Road and CrisisMappers and how they promote technology-supported civic participation in providing support for those in need post disasters. These studies use digital environments as a platform to promote volunteerism, and understanding the factors that build volunteer capacity online can provide key insights for organizers seeking to capitalize on post-disaster social support and convert this support into resources for the effected. Despite these findings, however, examinations of crowdsourced volunteer-capacity-building practices presented in Dittus et al. 2016 find deficiencies in crowdsourcing efforts in the aftermath of Typhoon Haiyan, suggesting post-disaster volunteer recruitment strategies are poorly understood. Relatedly, little study has sought to unify studies of these online volunteers and their roles in community recovery with processes established by governmental response organizations like FEMA and its ESF6 guidance.

Additionally, throughout these efforts, analyses of these volunteerism efforts are primarily facilitated by qualitative methods among few crises, leaving opportunities to study volunteerism and local community-based volunteer groups from a quantitative perspective and evaluate consistencies across disaster events. To this end, this paper concentrates on Twitter as a model platform for studying volunteerism discourse, as Twitter is an popular social media platform for discussions of disaster (see , e.g., Reuter, Backfried, et al. 2018 or Mccreadie et al. 2019). According to Petrovic et al. 2013 and C. Buntain et al. 2016, Twitter is a good place for finding volunteer-related content since it appears to have better coverage of the long-tail discussion that would come post crisis. Likewise, social media spaces like Twitter and Facebook are especially valuable for crowdfunding efforts (Borst et al. 2018; Lu et al. 2014), suggesting further understanding of volunteerism *on these platforms* can enhance conversions from observers of a crises to volunteers and donation resources. These studies therefore motivate us to study how to provide computational support in this emergency support function, which could help practitioners to accelerate and expand the horizon of the recovery process.

Crisis Informatics and Using Social Media to Improve Situational Awareness

While the above suggests much of the work on volunteerism discourse in social media spaces is qualitative in nature, a large volume of quantitative work exists on social media and disaster, specifically in the crisis informatics context. Much of this work, however, has tended to focus on the problem of locating tweets that contain crisis-relevant information during disasters as a means to improve situational awareness for disaster-management personnel (Mccreadie et al. 2019). Such studies mainly focus on identifying and retrieving actionable content (e.g., Purohit et al. 2018, Acerbo and Rossi 2017, Piscitelli et al. 2021, Rossi et al. 2018, and Longhini et al. 2017), which includes a broad range of information, from requests for search and rescue to messages of caution and advice. CrisisLex (Olteanu, Castillo, et al. 2014), for example, builds a lexicon of crisis-related terms that tend to frequently appear across various crisis situations with a focus on increasing recall in identifying crisis-relevant discourse. These same authors have also developed a dataset consisting of various disasters, called CrisisNLP (Imran et al. 2016), which gathers human-annotated crisis-related messages. TREC-IS similarly provides multiple Twitter datasets collected from a range of past wildfire, earthquake, flood, typhoon/hurricane, storm, bombing, COVID, tornado, explosion, fire, accident, hostage and shooting events manually annotated by expert response officers into 25 information types based on the information each tweet contains, such as "contains location" or "reports of emerging threats". While each of these studies contributes to our understanding of crisis-relevant information during mass emergency situations, their empirical findings focus on social media as an information source to identify actionable content, broadcast useful information and raise awareness in the immediate aftermath of an event. However, communities in recovery are also likely to rely on social media for information sharing, as we discuss above, but may not benefit from the same sort of situational awareness goals presented in the crisis informatics literature. In fact, a recent retrospective on the state of crisis informatics by Reuter, Backfried, et al. 2018 analyzes trends in crisis informatics literature, wherein they find significant focus on large events but limited impact for real-world disaster-management personnel. Rather, in a different retrospective of the field, Reuter et al. explicitly highlight needs for supporting citizen-to-citizen assistance and volunteering as an area of future work for crisis informatics (Reuter and Kaufhold 2018). Here, we aim to advance this area by evaluating consistencies in such discourse and developing methods for identifying specific instances of volunteerism messaging during crises as a path to make crisis informatics more useful beyond increasing situational awareness.

Text Mining and Domain Adaptation Across Crises

Transfer learning and domain adaptation are well-studied subjects in Natural Language processing. Transfer learning solved this problem of data deficit and poor model generalization by allowing us to take a pre-trained model of a task and use it for others. Promising results from previous works on different types of transfer learning for NLP include domain adaptation, cross-lingual learning, multi-task learning and sequential transfer learning (Pan and Yang 2009; Xia et al. 2015; Peters et al. 2018; Howard and Ruder 2018). These results motivate us to investigate transfering the knowledge of a pre-trained model on previous crisis data into a new unseen events.

We know at the time of any new crisis, there would be the millions of Twitter messages ("tweets") broadcast at any given time about that crisis and knowing what information to look for is often difficult. Annotating all these tweets is time-consuming which is not ideal since we need rapid response to new events. Also, the annotation task is costly, because we need expert annotators (the same process used in TREC-IS (Mccreadie et al. 2019) paper) to have high quality data. Although all mentioned studies in crisis informatics assumes sufficient consistency across different events and builds general models accordingly, the issue of differences across events and how well models generalize to new events and event-types is still a challenge as mentioned in the SMERP workshop report (S. Ghosh, K. Ghosh, Ganguly, et al. 2019). Olteanu, Vieweg, et al. 2015 shows differences in the distributions of information types across several disasters. To expect a decent performance model that applies an existing model to a new event, first, we need to evaluate a hypothesis that sufficient overlap exists in language across events. In C. L. Buntain and Lim 2018, the authors study how similar the lexicons are in response across disasters in online social network contents with the focus mostly on Twitter data. According to their study, commonalities emerge within similar disasters. Also, they gather words that are common across disaster types such as "victims"/"affected" and "prayer". However, what is important for our studies is how consistent the language around volunteer- and donation-related languages are across different events. By evaluating the consistency of language in volunteer-and donation-related social media content across different disasters and identifying the common pattern among them, we could assume there is enough similarity between events to train one super model that is transferable to the new and unseen events.

RESEARCH DESIGN

To identify applicability of volunteer- and donation-related social media messages across disasters of varying types, we conduct three experiments. Each study seeks to address one of the research questions. In the first experiment, we analyze consistencies in language about volunteerism and donations across disaster events and event types. This analysis tests our first research question if sufficient overlap exists in volunteerism content across events to make a cross-event model useful. Then to answer second research question, we train a simple machine learning model to measure the performance of this cross-event model on unseen events. We also evaluate some common sampling strategies like up- and down-sampling and compare them to tailored re-weighting strategies for crisis events of similar types. Results of the first and second studies suggest that sufficient overlap exists in the language of volunteerism that we can build standard machine learning models to identify useful content.

The third study answers whether domain experts' insights about the social context of crises increase the accuracy of standard machine learning models. We design the third experiment to test the impact of integrating domain experts' insights to identify available unlabeled data sources for expanding training data. To this end, we analyze which sources of data are useful for augmenting our initial dataset and how we could collect this additional data. Then, we experiment with different weak supervision approaches for integrating these unlabeled data with the hand-labeled data. Reducing manual assessment requirements and therefore time needed to assess this content can facilitate practitioners' integration of experts' insights. Such rapid assessment and adaption are crucial for tailoring existing models according to the emerging need of a new crisis.

STUDY 1: CONSISTENCY IN VOLUNTEERISM ACROSS CRISES

We conduct the first study to evaluate our first research question. We analyze consistencies in language about volunteerism and donations across disaster events and event types.

Datasets

For our study, we focus on the TREC-IS (Mccreadie et al. 2019) dataset since it contains a considerable number of volunteer- and donation-related tweets. We extract volunteer- and donation-related information types. For TREC-IS data, we collect all labeled tweets from 2018, 2019, 2020 and 2021A datasets. This data contains tweets related to 78 different crisis events. Relevant crises for TREC-IS include 13 natural and man-made event-types: wildfires, earthquakes, floods, typhoons/hurricanes, storms, bombings, shootings, explosions, tornadoes, accidents, fire,

hostages and COVID-19. TREC-IS tweets are categorized into 25 high-level information types. From these 25 categories, we only consider "CallToAction-Volunteer", "CallToAction-Donations", "Request-GoodServices", and "Report-Service Available" as volunteer- or donation-related labels. Other information-types are mostly related to general information or actionable content, which cover other types of non-recovery content. We further exclude tweets annotated as "Irrelevant" from our dataset and those tweets that contain fewer than four words. From 98,391 total labeled tweets, 4,063 include volunteer-related labels, which is a relatively small portion of the data (about 4.1%).

As a prepossessing step, we remove punctuation, hyperlinks (i.e., URLs), emojis, and stop words like "https" and "http" from each tweet's text. We also transform letters to lowercase, extract bigrams, lemmatize the remaining words, and only keep NOUNs, ADVERBSs, ADJECTIVEs, and VERBs.

Method

To evaluate how similar volunteer-related messages across events and event types are, we measure similarity in language between pairs of event-types, under the expectation that recovery- and volunteer-related content will be largely similar regardless of underlying event-type. We use cosine similarity as a similarity metric to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, the higher the cosine similarity. We calculate cosine similarly between a representative embedding vector from each pairs of event-types using sentence-embeddings generated from a pre-trained Sentence-BERT (Reimers and Gurevych 2019) language model. This representative embedding is comprised of sentence-embeddings averaged over all tweets in a particular event-type. In testing other embedding methods, we saw limited differences in results. We also try to confirm that volunteer-related content should have higher similarity across different event-types compared to other information categories by measuring similarity in language between pairs of event-types for all 24 information categories independently and taking the mean over the similarity scores for all event types and comparing how similar the language are across all event-types for each information category. We binarize this dataset into: **"All-Volunteer"**, containing tweets labeled as "CallToAction-Volunteer", "CallToAction-Donations", "Request-GoodServices" and "Report-Service Available", and **"All-Non-Volunteer"**, containing tweets labeled with remaining information categories. We expect tweets from the latter group should not be very similar to each other compared to volunteer tweets since recovery- and volunteer-related tweets come from the same topic-oriented selection process.

Results

Figure 2 shows the pair-wise similarities across volunteer-related content from different event-types. As one might expect, natural crises have high similarities with each other, for example, floods and typhoons have extremely high similarities. In a less obvious example, floods and typhoons also have a high similarity with earthquakes. Also, we observe that more anthropogenic crises like bombings and shootings have lower similarity in recovery- and volunteer-related content compared to natural crises. Bombings and shootings cannot be discounted completely, however, as they are still similar to volunteer-related content posted around wildfires. This result motivates us to seek some heuristic for sampling strategies based on the hierarchy of event-type (natural vs manmade).

Table 1 shows the average similarity score across all events for each information types. As we expected, "Donations" and "Volunteer" information-types have higher similarity scores compared to other contents while "Service Available" and "Good-Service" have slightly lower scores. As shown in the bottom section of the table, one could get a higher similarity score by grouping all volunteer content together compared to the score obtained from tweets from other content. This result is consistent with our expectation that non-volunteer tweets should not be very similar to each other compared to volunteer tweets since volunteer-related content comes from the same topic-oriented selection process. Figure 3 shows examples of volunteerism tweets. Even though these tweets were posted at different times and for different crises, we could still observe the consistent pattern in tweets' language asking for donations and other volunteer supports.

The results from this study answer our first research question. Therefore, we could assume there should be decent similarity in volunteerism and recovery language across different events and event-types. Based on these findings, we conduct our next study to evaluate if by leveraging the similarity in volunteer data, we could construct a general machine learning model that is transferable from one crisis to the other.

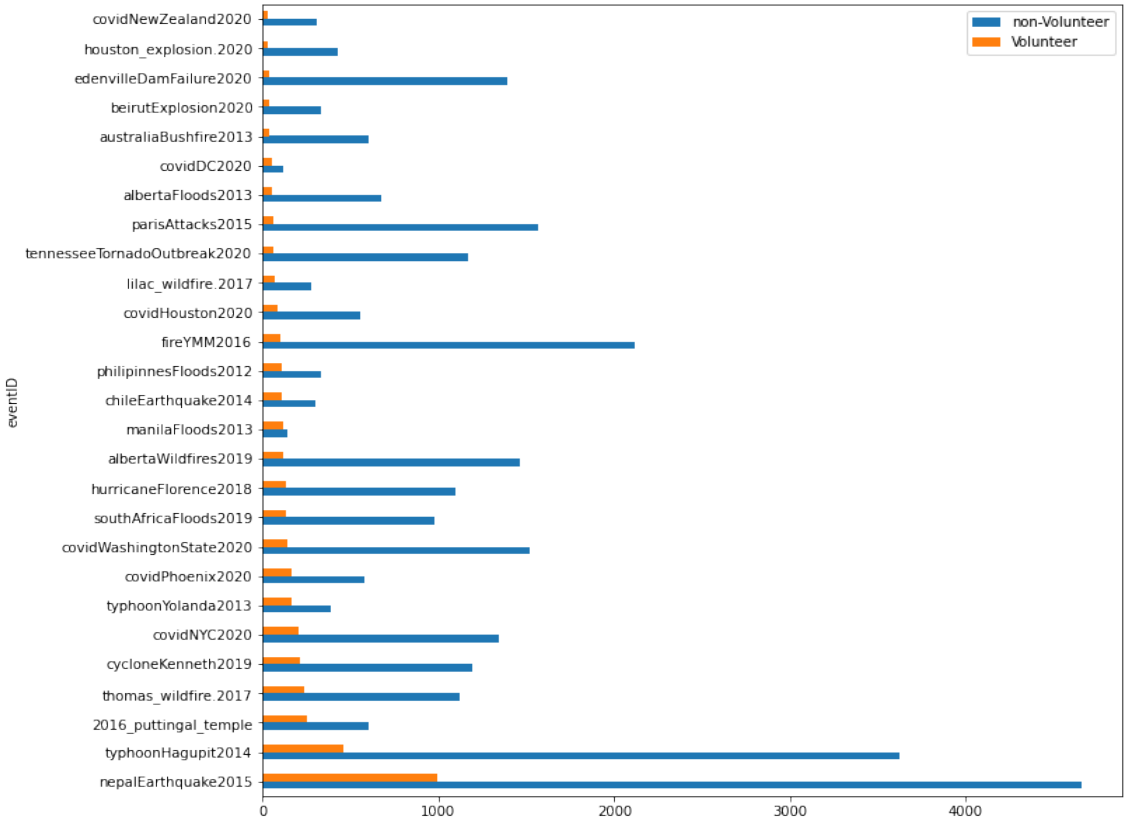


Figure 1. Distribution of recovery-related tweets over crises of evaluation set. A clear imbalance exists between recovery- and non-recovery-related data.

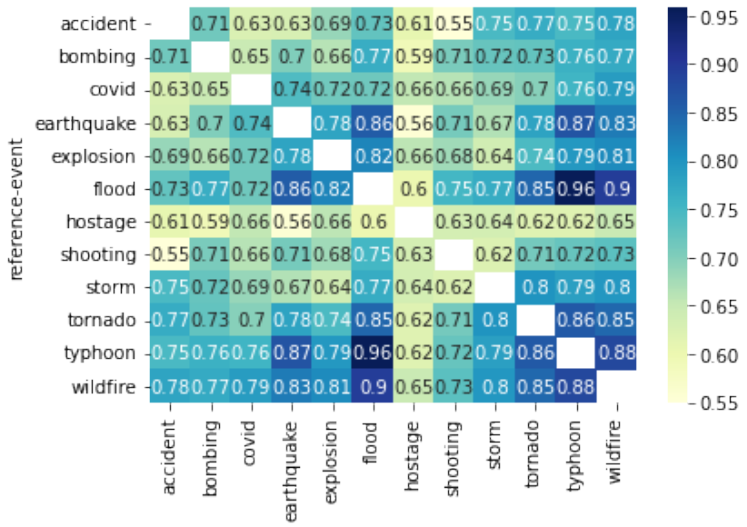


Figure 2. Heat map of event-types similarity on volunteerism tweets. The similarity matrix is obtained by generating sentence embeddings for each tweet using a pre-trained language model and taking the average for each event-type and measuring cosine similarity among them.

Information type	mean	std	# tweets	+95ci	-95ci
CallToAction-Donations	0.7425	0.1193	1099	0.7648	0.7202
CallToAction-Volunteer	0.6632	0.1519	290	0.6916	0.6348
CallToAction-MovePeople	0.6236	0.0983	882	0.6420	0.6052
Other-Advice	0.6703	0.1115	3781	0.6878	0.6528
Other-Sentiment	0.6552	0.1036	11627	0.6714	0.6389
Other-ContextualInformation	0.6422	0.1036	4884	0.6599	0.6245
Other-Discussion	0.6297	0.1036	5364	0.6460	0.6135
Report-Weather	0.6671	0.1236	8473	0.6956	0.6385
Report-MultimediaShare	0.6455	0.1076	24105	0.6624	0.6286
Report-News	0.6386	0.1095	19404	0.6558	0.6214
Report-Location	0.6332	0.1140	26120	0.6526	0.6137
Report-ServiceAvailable	0.6281	0.2179	2533	0.6653	0.5909
Report-EmergingThreats	0.6279	0.0970	7367	0.6460	0.6097
Report-Official	0.6276	0.1159	3106	0.6458	0.6094
Report-ThirdPartyObservation	0.6223	0.1095	19060	0.6395	0.6051
Report-Hashtags	0.6018	0.1286	17208	0.6220	0.5816
Report-Factoid	0.5992	0.1215	11006	0.6183	0.5802
Report-FirstPartyObservation	0.5981	0.1217	5496	0.6172	0.5790
Report-NewSubEvent	0.5823	0.1134	2919	0.6016	0.5629
Report-CleanUp	0.5806	0.1690	516	0.6121	0.5490
Report-OriginalEvent	0.5402	0.1175	5046	0.5602	0.5201
Request-GoodsServices	0.6037	0.2144	361	0.6532	0.5542
Request-InformationWanted	0.5414	0.1475	509	0.5645	0.5182
Request-SearchAndRescue	0.4876	0.1854	308	0.5304	0.4447
All-Volunteer	0.7270	0.0855	4063	0.7416	0.7124
All-Non-Volunteer	0.6417	0.1083	50927	0.6587	0.6247

Table 1. Similarity scores across different information categories average over different event-types from TREC Dataset. The top section shows the average similarity score for each information category. Volunteer-related information categories are shown in bold. "Donations" has the highest similarity score among all information types and "Volunteers" has a relatively high score. However, "ServiceAvaiable" and "GoodService" achieve lower scores. The bottom section shows the average similarity across all tweets annotated as any of the 4 volunteer-related labels and similarity score for tweets belong to other information categories. The first group getting higher similarity score justifies the validity of our first hypothesis that there is sufficient overlap exists in the language of volunteerism across events. First and second columns shows mean and standard deviation respectively average across all tweets labeled as the corresponding information type. Third column shows number of tweets for each information type. There are some overlaps since each tweets could be annotated by more than one category. The last two columns indicated the lower and upper bounds of the 95% confidence interval respectively.

STUDY 2: TRANSFERABILITY OF VOLUNTEERISM DATA ACROSS CRISES

In this study, we examine how we can build general classification models that are adopted across different events and event-types to answer our second research question. We also train standard machine learning models with different common sampling strategies and tailored re-weighting strategies and compare their performances.

Datasets

To extend our volunteer-related tweets, we look at and CrisisNLP (Imran et al. 2016) data, from 9 information types offered in Imran et al. 2016, we consider the “Donation needs or offers or volunteering services” type to identify requests and offers for volunteering. We collect 594 tweets from this set as well, adding these new tweets to the TREC dataset, raising the total number of volunteer-related tweets to 4,657.

As a prepossessing step, we remove punctuation, hyperlinks (i.e., URLs), emojis, and stop words like "https" and "http" from each tweet’s text. We also transform letters to lowercase, extract bigrams, lemmatize the remaining words, and only keep NOUNs, ADVERBSs, ADJECTIVEs, and VERBs.

Method

To assess the transferability of the model, we do cross-validation experiments using a simple support vector machine model (SVM). We use TF-IDF as a feature vector. We do experiments with different embedding methods such as BERT but we saw limited differences in result. Consequently, we decide to use SVM and TF-IDF to keep the model as simple as possible since the main focus of the study is how to choose the best sampling strategies and how and

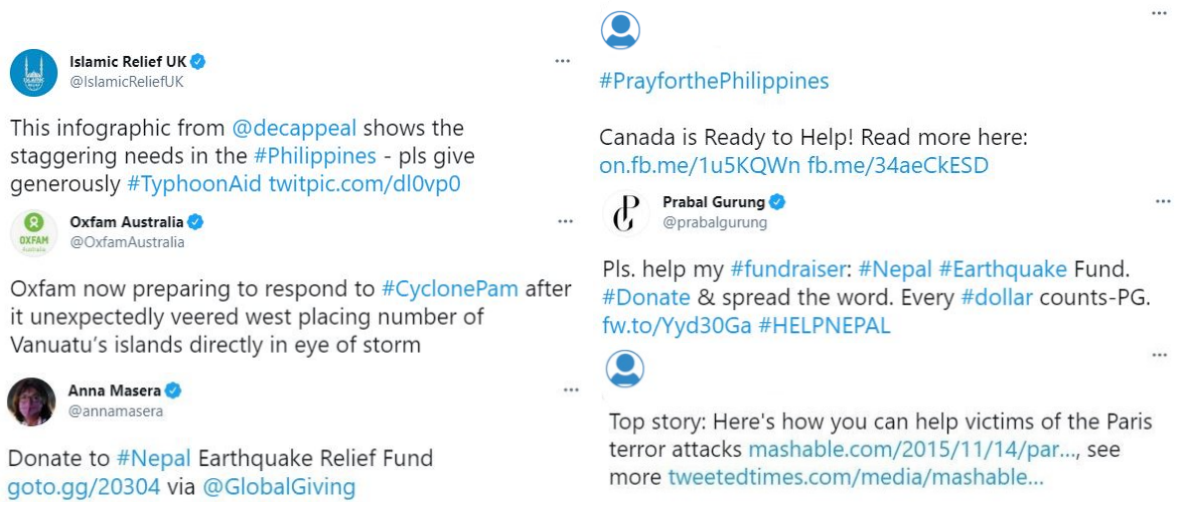


Figure 3. Examples of Volunteer- or Denotation-related tweets. In the first column, the first tweet is related to 2013 typhoon Yolanda, the second one is for 2015 Cyclone Pam and the last one is related to 2015 Nepal Earthquake. In the second column, the top tweet is for 2014 Philippines Typhoon Hagupi, the middle one is related to 2015 Nepal Earthquake and the bottom one is about donations for victims of 2015 Paris Attacks. Even though all of these tweets related to different events happened over different years, we could observe the similar language among them related to donations and volunteerism. Usernames are masked to protect the privacy of the users.

which data to integrate to mitigate the imbalance problem. Once decided o the best strategies, we could easily adapt more complicated models like neural network models to use them.

For the cross-validation, we hold out one event, train on other events and report the evaluation metrics on the held-out event. We use precision, recall, and F1 as our evaluation metrics. Our goal is to increase the F1 score since it shows a balance between precision and recall. Having high precision while maintaining high recall is important since we need to be able to collect more volunteer related tweets to cover different types of volunteerism contents (having high recall) while trying to lessen the number of false-positives (i.e., maintaining high precision).

The main challenge we face during training is how to lessen the effect of class imbalance issues, since the volunteer-related data is a small portion of our labeled data. For handling class imbalance problem, we experiment with different re-balancing strategies and compare their performance. We evaluate 8 different event-driven sampling and re-weighting strategies. We will discuss each of them briefly.

- **No Re-balancing without any Up-weighting (NONE)**: We use the imbalanced data that we have without balancing or re-weighting our data in any form.
- **No Re-balancing with Up-weighting (NONE-UPW)**: We use the imbalanced data but assign higher weights to the training data with similar event-type as a held-out event while training.
- **Down Sampling without Up-weighting (DOWN)**: We down-sample from the majority class, without replacement and we assign the same weight to all the training data.
- **Down Sampling with Up-weighting (DOWN-UPW)**: We down-sample the majority class, but this time we assign higher weights to the samples from the same event-type of a held-out event.
- **Up Sampling without Up-weighting (UP)**: We repeatedly take samples, with replacement, from the minority class until the class is the same size as the majority with equal weight for all samples.
- **Up Sampling with Up-weighting (UP-UPW)**: We up-sample the minority class, but this time we assign higher weights to the samples from the same event-type of a held-out event.
- **Event-type weighting scheme (UP-EV)**: In this sampling strategy, during up-sampling minority class, events of the same type are 10 times more likely to be sampled. But equal weight is assigned to all the training data during training.

- **Hierarchy of event-types (UP-HEURISTIC)**: We categorize our event-types into two categories: **man-made** like shooting, **natural** such as earthquake and **general** when using other source of unlabeled information for data augmentation. We annotate data of the same event-type with the highest weight, data of the same "kind" of the event (manmade vs. natural) weighted 6, annotated data of other "kinds" of events weighted 3, and augmented data with the lowest weight. We sample data according to these weights.

Improving Estimates of Variation To evaluate how our model performs across different unseen events, we hold out one event as a test and train our SVM model on the remaining data. One of the problems in this approach is that the volunteer data is not normally distributed among all events. For example, events related to earthquakes contain a much larger share of volunteer-related tweets than shooting-related events. For this reason, we exclude individual events with fewer than 30 examples of volunteer tweets from our evaluation set. We end up with only 27 events as our evaluation set. As shown in Fig 1, this volunteer-related content is exceedingly rare. To estimate variation across this small number of events which we can use as a evaluation set, we use re-sampling with replacement to create new training sets for each of the 27 events we use for evaluation following the approach proposed by Beleites and Salzer 2008 to increase the robustness and confidence of our result. We have $k = 78$ events in total, and $n = 27$ events where we have sufficient positive samples for estimation, where n « k. We want to replicate each event-fold z times to boost the number of observations we have for our models' performance metrics, which should provide $n \times z$ total estimates of precision, recall, and F1. We estimated z should be equal to 20 for achieving 0.025 standard error, which sufficiently constrains standard error measures such we obtain tighter confidence intervals on model performance relative to the variation across methods (i.e., we want to avoid wide confidence intervals that are driven solely by the small number of events). Also, since this procedure is computationally intense, we run an evaluation for every single re-sampled fold for a given event on a separate cluster in parallel and aggregate the result.

Results

Table 2 shows different sampling and weighting strategies for different performance metrics. While "NONE" achieves the highest precision score, all up-sampling strategies also get comparably high precision. Both down-sampling strategies get lower precision scores compared to other sampling and weighting strategies while achieving the highest recall. Most up-sampling strategies also maintain good recall scores. Therefore, as shown in table 2, "UP", "UP-UPW", "EV-Up" and "UP-HEURISTIC" achieve the highest F1-score with a very slight difference. All these 4 strategies outperform other weighting and sampling strategies on recall while preserving relatively high precision and therefore high F1 score.

method	precision	recall	f1_score	f1_ci	pr_ci	re_ci
UP-EV	0.46362227331499767	0.6483887220908487	**0.5196421615164815**	0.014840919874698611	0.01527648071838506	0.015118554133924907
UP	0.44484703238710377	0.66954601348409	0.5151524517303945	0.0148156079269787	0.015223823083567744	0.015020806721141958
UP-HEURISTIC	0.45525515273745476	0.6489296063325118	0.5131033246585379	0.014787955666297667	0.015390106048238145	0.015041059572206288
DOWN	0.3854112989249049	**0.7132698157192792**	0.4780747939643004	0.014719894130712497	0.014402532369603713	0.014714083379137686
UP-UPW	0.41487700255178805	0.5103035188741323	0.4384878234896081	0.013279318198064558	0.015373907611752925	0.013506650386880375
NONE-UPW	0.47955414179096717	0.4036340324515155	0.42081898584360616	0.01256219474918321	0.015452133212385921	0.01261356073371204
DOWN-UPW	0.3175070623498344	0.6833994377123236	0.4120168005623443	0.014325301533015599	0.014304293802491479	0.011974368465014283
NONE	**0.7209994535977459**	0.20344515707743355	0.28928005927750894	0.014678212669937183	0.01769623595968339	0.01293471647714322

Table 2. Performance of our Model for different sampling strategies sorted by F1 score. All four variations of up-sampling strategy outperforms other re-weighting approaches on recall and F1-score while preserving relatively high precision.

STUDY 3: WEAK SUPERVISION FOR INTEGRATING DOMAIN EXPERTISE

Previous studies (e.g., Ratner et al. 2019, Alfonseca et al. 2012, Bunescu and Mooney 2007, Mintz et al. 2009, Rekatsinas et al. 2017, Zhang et al. 2017) shed light on the utility of weak supervision and heuristic methods for many tasks. Similarly, practitioners are increasingly turning to weak supervision to reduce costs of manual labeling, especially when domain expertise is required. The existence of common patterns across crisis events; the sparsity of volunteer-specific manually labeled content; the cost of annotation; and the availability of massive, unlabeled data likewise leads us to experiments in using weak supervision for improving volunteerism classification systems. This ability could enhance computational assistance for emergency response officers and their core support functions by providing fast, low-cost paths to integrate new knowledge into our models. These potential enhancements stem from two sources: First, in times of crisis, the ability to respond fast is essential, but high-quality annotation is often slow, as new messages need to be labeled by costly expert annotators. Second, we have many useful pieces of data, other than tweets related to previous crises, that could reduce dependence on perfectly and fully labeled data. For example, we could use domain experts' knowledge to identify accounts that have high volumes of relevant content and integrate them with our initial datasets.

To examine these approaches, we pose the following research question: What source of weakly supervised volunteerism data leads to the largest performance improvements in our models? We therefore compare weakly supervised methods applied across three data sources: fully random Twitter data, a collection of unlabeled crisis-related content, and a dataset of tweets collected from a small set of aid organizations' Twitter profiles. These sources represent increasing levels of domain knowledge, from absolutely no domain knowledge needed to knowledge about domain-relevant search terms to deep insights about good sources for volunteerism content. Our expectation here is that these three sources will provide increasingly more samples using weak supervision, which will improve our models' performance. These experiments test different ways to identify and collect such data, and based on our results, we suggest optimal strategies for applying weak supervision to insights from domain experts.

Datasets

As our first weak-supervision data source, we construct a dataset of approximately 60k randomly sampled English-language tweets. This sample is drawn from a collection of complete timelines for 5,000 US Twitter users developed at the Center for Social Media and Politics, where US accounts have been geolocated via network analysis. Each user should have an approximately complete timeline up to the end of this archive in 2019 and is restricted to users who posted more than 100 times between January 1, 2015 and December 31, 2017. We then randomly select tweets from across these users' timelines, which we call the "Random" dataset.

Our second source of weakly supervised data comes from a large pool of unlabeled crisis-related content from TREC-IS. While the first source can contain any kind of social media content, this second source is focused on ostensibly crisis-related content that TREC-IS organizers collected using crisis-related search terms, similar to the CrisisLex collections Olteanu, Castillo, et al. 2014. From this set, we randomly select 59,734 tweets from unlabeled TREC-IS data and call this dataset the "UnlabeledCrisis" data.

Lastly, to augment our data with experts' domain knowledge, we extract the top-5 most common volunteer organizations mentioned in labeled tweets, which include UNICEF, Red Cross (America, Canada, and Philippines), and World Food Program. We further extend this list using the volunteer organizations mentioned in ESF6 (FEMA 2016). These accounts represent a set of relevant sources for volunteerism information and are entities with which practitioners are already familiar (as the ESF6 list demonstrates). Consequently, a domain expert might be able to construct such a list quickly and far more rapidly than actual message annotation. Pulling the 3,250 most recent tweets from accounts associated with these organizations (Table 3), we construct a new dataset, referred to as the "NGO" dataset.

NGO Twitter Account	Collected Tweets	Description
American Red Cross	3250	The American Red Cross is a non-profit humanitarian organization provides emergency assistance and disaster relief in the United States.
Canadian Red Cross	3250	The Canadian Red Cross provides assistance to Canadians experiencing an emergency or disaster.
UNICEF	3250	UNICEF is a United Nations agency responsible for providing humanitarian and developmental aid to children worldwide.
Philippine Red Cross	3250	The Philippine Red Cross is committed to provide quality life-saving services especially for indigent Filipinos in vulnerable situations.
World Food Programme	3250	The World Food Programme is the food-assistance branch of the United Nations.
Operation Blessing Foundation Philippines, Inc.	3250	Operation Blessing is a non-governmental organization accredited by the Philippine Council for NGO Certification as a donee institution.
Points of Light Foundation and Volunteer Center National Network	3243	Coordinates unaffiliated volunteers and meets the needs of the local community and other disaster response agencies.
National Voluntary Organizations Active in Disaste(NVOAD)	3165	NVOAD is a nationwide coalition of organizations that work together in all phases of disaster.
Jewish Response to Disaster	3050	NECHAMA is a volunteer-driven nonprofit headquartered in the Twin Cities of Minnesota.
MNA TAG Disaster	619	Presbyterian Church in America (PCA) Mission to North America (MNA) Disaster

Table 3. NGO accounts that appeared in volunteerism-labeled data or are mentioned in FEMA. We pull the most recent 3250 tweets from these volunteer organizations' tweeter accounts. The difference in the number of fetched tweets is because some organizations posted fewer than 3250 tweets in total.

Method

Given these three data sources of increasingly specific, we answer our research question by evaluating modeling performance after applying weak supervision methods to these sources. To this end, we apply several weak

supervision approaches to this data to create new, augmented training sets. For consistency, we run all these evaluations by retraining the best-performing model from above, UP-EV, on the augmented datasets.

In our first approach, we augment our training set with the NGO dataset, naively labeling all content from NGOs as being positive samples, or instances of volunteerism. This approach represents a basic for of weak supervision that has little technical requirement or sophistication but may nonetheless perform well, as these NGO sources primarily share recovery-related content.

In second setting, we use the label spreading (Zhou et al. 2004) algorithm to annotate unlabeled data. This algorithm "spreads" information from labeled points to unlabeled points based on the similarity between labeled and unlabeled samples. E.g., an unlabeled message that is most similar to several volunteerism-related messages will have that label "spread" to it. As the feature vector for the label spreading algorithm, we use word-embeddings, similar to Study 1, and have evaluated several embedding methods for robustness, including "Glove Tweeter 100", "Glove Tweeter 200" and "Fast-text". We see limited variation in performance across embedding models, suggesting our results are robust to embedding selection; got forward, we use "Fast-text" as it appears to have the best performance. Therefore, we report our result with "Fast-text" as a word-embeddings and 0.8 as a threshold for the rest of the paper. Finally, we augment the original training set with all tweets that receives the "volunteerism" label via label spreading and retrain our models.

Lastly, our third approach uses semi-supervised methods. We train a logistic regression model using our labeled data with word-embeddings as features and add those samples from the unlabeled data that receive a high probability (> 0.8) of being a volunteerism sample to the original training dataset. We have experimented with different thresholds and found that a threshold of 0.8 outperforms others in both precision and recall.

In summary, we retrain and evaluate the performance of UP-EV model with 9 different settings regarding the augmented data and weak supervision approaches:

- **Baseline** UP-EV
- **Augmented by Random Data** Labeled via semi-supervision (Semi-Random+Labeled)
- **Augmented by Unlabeled Crisis Data** Labeled via both label-spreading (LS-UnlabeledCrisis+Labeled) and semi-supervision (Semi-UnlabeledCrisis+Labeled)
- **Augmented by NGO Data** All naively labeled as "volunteer" (All-volunteer-NGO+Labeled)
- **Augmented by NGO Data** Labeled via both label-spreading (LS-NGO+Labeled) and semi-supervision (Semi-NGO+Labeled)
- **Augmented by NGO and Unlabeled Crisis Data** Labeled via both label-spreading (LS-NGO+UnlabeledCrisis +Labeled) and semi-supervision (Semi- NGO+UnlabeledCrisis +Labeled)

Results

Table 4 indicates that labeled data augmented by NGO and annotated via semi-supervision outperforms each sampling and weak-supervision strategy on F1 and precision, while preserving relatively high recall. We could observe the same trend for labeled data augmented by both NGO and UnlabeledCrisis data, and annotated by a semi-supervised approach. Highest recall is related to labeled data augmented by NGO and UnlabeledCrisis data, and annotated by Label-Spreading. Table 4 also illustrates the semi-supervised method (with any data setting) outperforms Label-Spreading and naive all-volunteer methods. Similarly, both label-spreading and semi-supervised approaches achieve higher scores when augmenting our labeled data with both NGO and UnlabeledCrisis data compared to the baseline 'UP-EV' method without weak supervision.

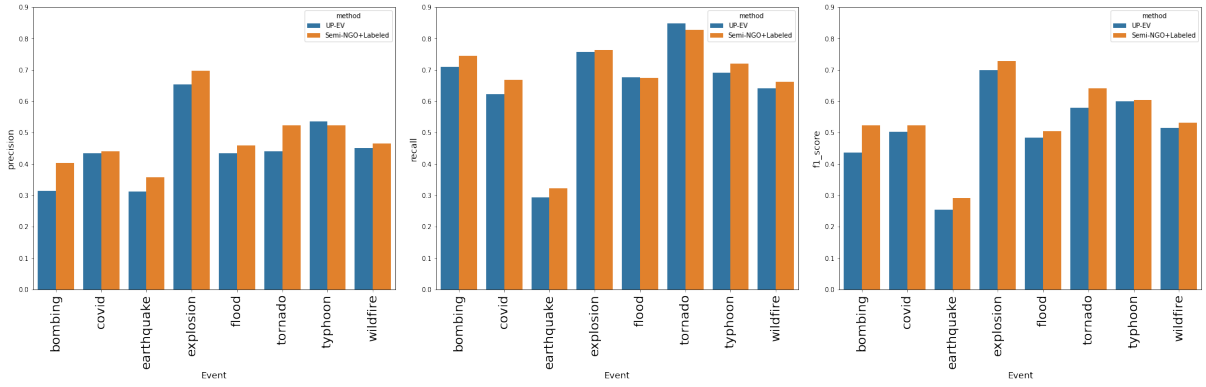
These results highlight two points: First, they show the utility of weak supervision for the volunteerism domain, as augmenting our training data with weakly supervised labels increases overall performance. Second, these results inform the the importance of integrating domain knowledge to identify additional sources of data to augment the initial dataset. Surprisingly, we could observe limited difference between the performance of UnlabeledCrisis and Random data despite our expectation that relevant crisis data would prove more useful than a truly random dataset, which may contain large volumes of pop-culture and entertainment references. In fact, augmenting with Random data achieves slightly better recall than UnlabeledCrisis data, though, less surprisingly, in both precision and F1, UnlabeledCrisis data shows slightly better performance compared to Random data.

In Figure 4 and 5, we study the effect of integrating the dataset with additional unlabeled crisis corpus identified by domain experts for each event-type and event respectively. These comparisons test whether the results are

consistent with what we have seen in general trends (averaging over all event-types or events). We only report result for eight of the thirteen event-types (others are omitted for too little data), including bombings, COVID, earthquakes, explosions, floods, tornadoes, typhoons and wildfires. Figure 4 shows all event-types benefit from integrating domain knowledge with regard to F1 score since all event-types get higher F1 scores using model trained on "Semi-NGO-Labeled" data. Regarding precision, all event-types except typhoons (which gets slightly lower score compared to "UP-EV") achieve higher precision using "Semi-NGO-Labeled" data. Additionally, almost all event-types benefit from integrating other sources of data during training process for recall. The only exceptions are floods and tornados, both of which achieve scores comparable to models trained only on hand-labeled data. Figure 5 shows detailed comparison between two models, "UP-EV" and "Semi-NGO+Labeled", for each event.

method	precision	recall	f1_score	f1_ci	pr_ci	re_ci
Semi-NGO+Labeled	**0.484690326510111**	0.670015482236968	**0.5430757379491981**	0.014642617061920162	0.015133022691270542	0.014838692826945513
Semi-NGO+UnlabelledCrisis+Labeled	0.47618814485323074	0.6694252257890847	0.5364832570046804	0.015225021420742755	0.016093096524792287	0.014658189126596844
Semi-UnlabelledCrisis+Labeled	0.45577122315508994	0.677959812431114	0.5245908201144224	0.014517073694202092	0.01509047122564312	0.014782582348527317
Semi_Random+Lableled	0.44540396723631875	0.6899158039402074	0.5206537742228666	0.014411670561331025	0.014807907819918155	0.01533092302236643
UP-EV	0.46362227331499767	0.6483887220908487	0.5196421615164815	0.014840919874698611	0.01527648071838506	0.015118554133924907
LS-NGO+Labeled	0.4358083023961947	0.6998034255632005	0.5167804948650097	0.015068327070856534	0.015273322190317642	0.014953710375501363
LS-NGO+UnlabelledCrisis+Labeled	0.3896762868851393	**0.7312302737969972**	0.4847963658173305	0.01570155149184068 7	0.01592885949399968	0.01485845672318689
LS-UnlabelledCrisis+Labeled	0.38935472678782057	0.7202094463539347	0.48059522658567627	0.014988211260089049	0.015188688043551678	0.014502122064445917
All-volunteer-NGO+Labeled	0.3741964594202213	0.7073406585423956	0.46149239762328925	0.015037365433277564	0.015098241418655053	0.015121041905554822

Table 4. Model Performance Across Weak-Supervision Strategies, Sorted by F1 Score. In general, integrating domain expertise appears to improve performance in all three metrics. Semi-supervised approaches outperform other integration strategies.



(a) Precision **(b) Recall** **(c) F1 Score**

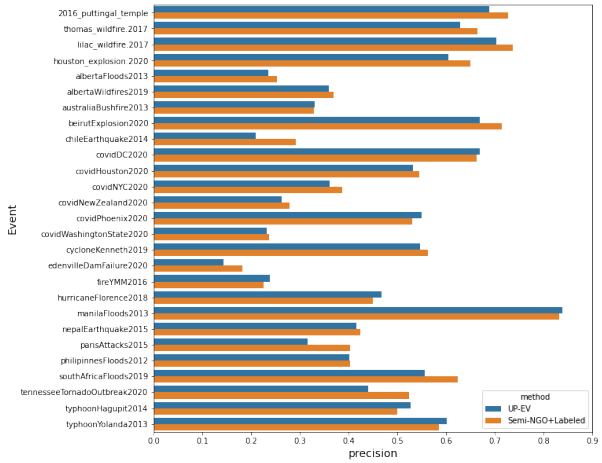
Figure 4. Cross-Event Performance for Baseline (UP-EV) and Best Weak-Supervision Models ("Semi-NGO+Labeled"). All event-types appear to benefit from added domain expertise to augment training data.

POST-STUDY ANALYSIS

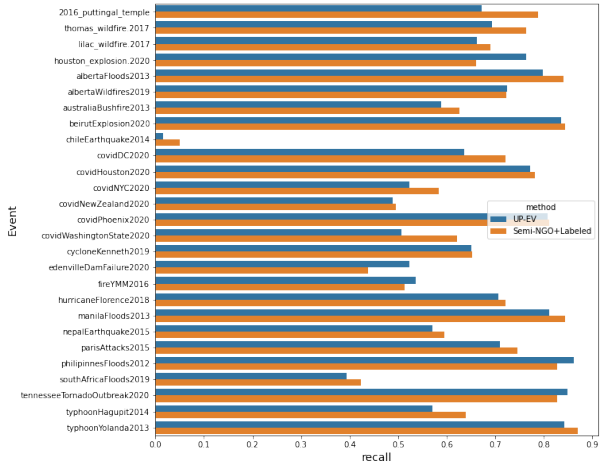
Table 4 indicates labeled data augmented either by Random or by UnlabeledCrisis gets comparatively the same performance in all three metrics. This finding contradicts our expectations, as random content in social media is much less likely to contain relevant crisis information than the unlabeled crisis dataset. To evaluate why there is such limited difference between the Random and UnlabeledCrisis sets, we examine the resulting set of tweets that get added to our augmented datasets from all three different sources.

Surprisingly, with semi-supervision using the EV-UP model and an inclusion threshold of 0.8, we find only 1,753 and 1,771 tweets are added from the Random and UnlabeledCrisis datasets respectively. In contrast, this same procedure produces 6,051 weakly supervised volunteer-related tweets from the NGO dataset. Contrary to our expectations, little difference exists in relevant but unlabeled data between the Random and UnlabeledCrisis datasets.

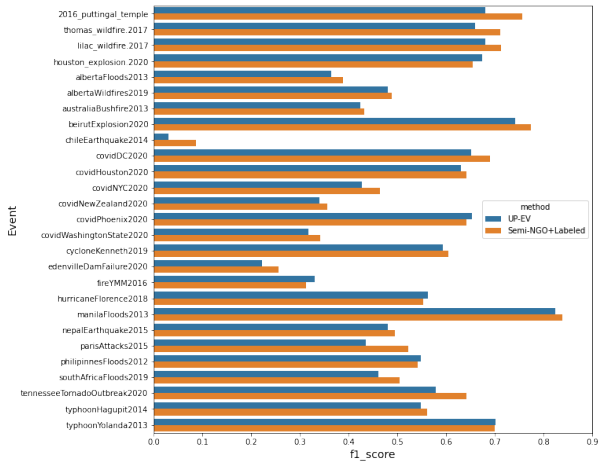
One possible reason is class imbalance issues for volunteer content. UnlabeledCrisis would not be as effective as NGO data since the main focus is not volunteer-related content and we know from Study 1 that there is much less volunteer-content compared to other information-types. The small number of retrieved positive tweets from UnlabeledCrisis might be because the volunteer-related data is a small portion of this data. NGO dataset contains more volunteer- and donation-related content since the NGO dataset is the only source in our experiment that is gathered using expert knowledge. In contrast, the UnlabeledCrisis dataset contains crisis data without considering the types of message the tweet contains (volunteer or not). This difference might explain why UnlabeledCrisis data has almost the same performance as Random dataset. Therefore, we could conclude, when identifying external



(a) Precision



(b) recall



(c) F1 Score

Figure 5. Comparison the performance of two models : "UP-EV", which applied "Event-type weighting scheme" on only labeled data; and "Semi-NGO+Labeled", which apply "UP-EV" sampling strategy on labeled data augmented with NGO data, aggregating the event-level performance. Though we could see 5 out of 27 events, namely fireYMM2016, hurricaneFlorence2018, manilaFloods2013, philiphinesFloods2012 and typhoonYolanda2013, get lower F1 score when trained on augmented data, the majority of events benefit from integrating domain's expertise to expand the initial seed in F1 metric.

sources, it is more effective to find more volunteer-related sources (such as NGO's accounts) rather than using any large available unlabeled crisis corpus. This recommendation could be even more important when we are dealing with types of data that are rare compared to other information types.

DISCUSSION

In this work, we seek to improve on previous works on volunteerism discourse in social media spaces (which are mostly qualitative in nature) by introducing the quantitative and computational approach that could merge massive available data posted on social medias during and aftermath of the crisis with domain expertise to help emergency response officers and stakeholders looking to social media for coordinating volunteers and donations.

On The Applications of Weak supervision to Other Information Types

We focus on volunteerism because it has not been sufficiently studied in the post-disaster case, but such approaches could be used for other information types as well. Our methodologies can be applied to any information type as far as there is enough consistency across varying events. Applying the same pipeline and using Table 1 as a guide reference, one could evaluate the existence of overlap in language across different events and implement a generalized and pre-trained model that is transferable across different events. The method we apply above for estimating standard errors could be useful to deal with the limited number of volunteerism data and increase the confidence of the reported result by creating multiple random copies of the available data. We believe this approach could be generalized to other domains whenever facing the similar issue of limited available observations. It may be a useful method to apply by future participants of the TREC-IS track to increase the confidence of their results and implement more robust models. Using other techniques like few-shot learning (Wang et al. 2020; Sun et al. 2019) and adaptors (Pfeiffer et al. 2020) could be applied to deal with few available volunteerism contents by leveraging a large number of similar tasks in order to learn how to adapt a base-learner to a new task for which only a few labeled samples are available if using more complicated models, for example, Recurrent Neural networks or Transformer-based models. But it is out of the scope of this paper, since our main focus in on how to integrate the available data using domain experts' knowledge to get better results. We leave this as a future work.

Integrating Domain Experts' Insights

Knowledge of the domain and its stakeholders, and empowering domain experts to identify useful additional sources of data that could be integrated with initial hand-labeled dataset can enhance accuracy of the model to identify recovery and volunteerism content across various events. We integrate recent posts related to the top mentioned accounts in our labeled data along with a subset of volunteer organizations identified by ESF6 and unlabeled random data from the previous crisis and we have seen improvement in the accuracy of the machine learning models. There are many other resources that could be used namely replies(discussion) to the annotated tweets or tweets related to followers of the Top NGOs or frequently mentioned users in the annotated set. We could leverage Domain Experts' Insights into which data to include to expand our data-set without the need to annotate these data. It could be an interesting future research path to examine how we could leverage domain knowledge for other information types or other domains and how we could collect additional data that are useful to the goal that we want to achieve. For example, if the goal is to train models tailored for COVID-19 volunteerism, it may be useful to collect tweets from CDC or other health officials' Twitter accounts. In another scenario, if we want to train a model that could identify "Search and rescue" tweets for pets, we may collect tweets posted by pet-advocates communities across different social media's platforms. For each model we could get advice from the domain experts on how to expand our initial sets and make them more adaptable to the goal in mind.

Computer-Supported Cooperative Work in Engaging Local Voluntary Organizations

ESF6 has a top-level function on coordinating local voluntary and faith-based organizations in the recovery stage of a disaster, and while emergency response officers are likely to have know a core set of community relief organizations, local agencies may be unknown to them. Identifying these local groups is non-trivial task that may need lots of efforts. By leverage the knowledge from prior events, one could identify these voluntary organizations in an efficient and computational approach since we expect to be a consistency in the language of these groups as we have seen in the examples shown in 3. For example, the language for asking donation is mostly similar in both earthquake and flood events. Therefore, by having the tweets posted by local volunteer groups during earthquake, we have a higher chance to retrieve the new emerging faith-based organizations related to flood. Also, the mechanism that we have described for augmenting our data specially the NGOs-expansion could be helpful for this end. One could augment

the data with posts of a core set of known community relief organizations to identify local agencies since both of them using the similar language for coordinating volunteer supports and donation.

Also, by using volunteer organizations mentioned by FEMA, we limit the augmented data to US-based volunteer groups. It may be helpful to expand the list of NGOs with the list of volunteer groups identified by the National Emergency Support agencies for each country to get a broader knowledge about local voluntary organizations worldwide and for each region.

LIMITATIONS

Our first limitation is our model is reliant on extant labels of non-experts from TREC-IS. Even though we apply different weak supervision approaches to expand our dataset, there is still room to analyze how to make the model less labor-intensive. Also, since we only rely on data from TREC-IS and CrisisNLP, the trained model may suffer from the potentially incomplete view of an event. We may miss specific aspects of volunteer tweets that are not considered in information-types suggested by TREC-IS and CrisisNLP. Moreover, the way that we have expand our data may suffer some deficiencies. First, we do not know how much local insight we get from the data, especially with our NGO-based expansion. For example, we may miss local organizations, such as churches, that are opening up for shelter or food. It would be an interesting avenue for future work to try to expand the dataset specifically for each event based on the the main goal of the model (for example, focusing on food donation) by applying expert knowledge instead of using more broader approaches like what we have done with top NGOs that are applicable across all events.

Our work is also necessarily based on the language of the message, not the user, so we could miss content from NGO- or volunteer-type accounts that are relevant but use different language. Finally, our model is language-specific in evaluation, since our TREC-IS and CrisisNLP data set only contain English tweets. We may miss some valuable content that are posted in other languages than English, which is a known and common limitation among crisis-informatics systems (despite practitioners' requests for more multi-lingual solutions). If a crisis happens in the regions which speaking another language than English, it is highly probable that people post their volunteer-related contents in the native language since it is faster and easier to broadcast information, promote volunteering opportunities and encourage more local people to engage. Focusing on languages other than English could be helpful in expanding our dataset around volunteerism especially for local non-native English-speaking regions.

CONCLUSION

This paper outlines methods for providing computational support in this emergency support function by evaluating the consistency of language in volunteer- and donation-related social media content and developing semi-automated models for classifying volunteer- and donation-related social media content in new disaster events. It also evaluates how integrating experts' insights from emergency response officers could help to identify additional sources of data to get cheaper sources of labels and augment the initial dataset.

Results show volunteer- and donation-related social media content is sufficiently similar across disaster events and disaster types to warrant transferring models across disasters. Our study also sheds light on the importance of strategies for identifying external sources of data to be used for a weak-supervision approach by incorporating crisis-response domain knowledge into the learning task rather than just using available crisis corpus, especially when dealing with highly imbalanced data. These sources, if identified correctly, could improve classification accuracy and accelerate the emergency support in new events without the need for a costly and time-consuming annotation process. We hope this paper could help to better support emergency response officers in their core support functions around recovery and coordinating mass for rapid response to new events.

REFERENCES

Acerbo, F. S. and Rossi, C. (2017). "Filtering Informative Tweets during Emergencies: A Machine Learning Approach". In: *Proceedings of the First CoNEXT Workshop on ICT Tools for Emergency Networks and DisastEr Relief*. I-TENDER '17. Incheon, Republic of Korea: Association for Computing Machinery, pp. 1–6.

Alfonseca, E., Filippova, K., Delort, J.-Y., and Garrido, G. (2012). "Pattern Learning for Relation Extraction with Hierarchical Topic Models". In.

Beleites, C. and Salzer, R. (Apr. 2008). "Assessing and improving the stability of chemometric models in small sample size situations". In: *Analytical and bioanalytical chemistry* 390, pp. 1261–71.

Borst, I., Moser, C., and Ferguson, J. (2018). "From friendfunding to crowdfunding: Relevance of relationships, social media, and platform activities to crowdfunding performance". In: *New Media & Society* 20.4, pp. 1396–1414. eprint: https://doi.org/10.1177/1461444817694599.

Bunescu, R. and Mooney, R. (2007). "Learning to extract relations from the web using minimal supervision". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 576–583.

Buntain, C., Golbeck, J., Liu, B., and LaFree, G. (Mar. 2016). "Evaluating Public Response to the Boston Marathon Bombing and Other Acts of Terrorism through Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media* 10.1.

Buntain, C. L. and Lim, J. K. R. (Nov. 2018). "#pray4victims: Consistencies in Response to Disaster on Twitter". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW.

Cobb, C., McCarthy, T., Perkins, A., Bharadwaj, A., Comis, J., Do, B., and Starbird, K. (2014). "Designing for the Deluge: Understanding and Supporting the Distributed, Collaborative Work of Crisis Volunteers". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '14. Baltimore, Maryland, USA: Association for Computing Machinery, pp. 888–899.

Dittus, M., Quattrone, G., and Capra, L. (2016). "Analysing Volunteer Engagement in Humanitarian Mapping: Building Contributor Communities at Large Scale". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. CSCW '16. San Francisco, California, USA: Association for Computing Machinery, pp. 108–118.

FEMA (June 2016). *Emergency Support Function #6 – Mass Care, Emergency Assistance, Temporary Housing, and Human Services Annex*. Tech. rep. Federal Emergency Management Agency.

Ghosh, S., Ghosh, K., Chakraborty, T., Ganguly, D., Jones, G. J. F., and Moens, M., eds. (2017). *Proceedings of the First International Workshop on Exploitation of Social Media for Emergency Relief and Preparedness co-located with European Conference on Information Retrieval, SMERP@ECIR 2017, Aberdeen, UK, April 9, 2017*. Vol. 1832. CEUR Workshop Proceedings. CEUR-WS.org.

Ghosh, S., Ghosh, K., Ganguly, D., Chakraborty, T., Jones, G. J. F., and Moens, M.-F. (Jan. 2019). "Report on the Second Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2018) at the Web Conference (WWW) 2018". In: *SIGIR Forum* 52.2, pp. 163–168.

Glasgow, K., Fink, C., Vitak, J., and Tausczik, Y. (2016). "'Our hearts go out': Social support and gratitude after disaster". In: *Proceedings - 2016 IEEE 2nd International Conference on Collaboration and Internet Computing, IEEE CIC 2016*, pp. 463–469.

Howard, J. and Ruder, S. (2018). *Universal Language Model Fine-tuning for Text Classification*. arXiv: 1801.06146 [cs.CL].

Imran, M., Mitra, P., and Castillo, C. (2016). "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *CoRR* abs/1605.05894. arXiv: 1605.05894.

Longhini, J., Rossi, C., Casetti, C., and Angaramo, F. (2017). "A language-agnostic approach to exact informative tweets during emergency situations". In: *2017 IEEE International Conference on Big Data (Big Data)*, pp. 3739–3475.

Lu, C.-T., Xie, S., Kong, X., and Yu, P. S. (2014). "Inferring the Impacts of Social Media on Crowdfunding". In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. WSDM '14. New York, New York, USA: Association for Computing Machinery, pp. 573–582.

McCreadie, R., Buntain, C., and Soboroff, I. (2020). "Incident Streams 2019 : Actionable Insights and How to Find Them". In: *Proceedings of the 17th International Conference on Information Systems for Crisis Response And Management*. May.

Mccreadie, R., Buntain, C., and Soboroff, I. (Sept. 2019). *TREC Incident Streams: Finding Actionable Information on Social Media*. Ed. by Z. Franco, J. González, and J. Canós.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). "Distant supervision for relation extraction without labeled data". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). "Crisislex: A lexicon for collecting and filtering microblogged communications in crises". In: *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14)*. CONF.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to Expect When the Unexpected Happens: Social Media Communications Across Crises". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '15. Vancouver, BC, Canada: Association for Computing Machinery, pp. 994–1009.

Pan, S. J. and Yang, Q. (2009). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). "Deep contextualized word representations". In: *Proc. of NAACL*.

Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., and Shrimpton, L. (June 2013). "Can Twitter Replace Newswire for Breaking News?" In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. (2020). "AdapterHub: A Framework for Adapting Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 46–54.

Piscitelli, S., Arnaudo, E., and Rossi, C. (2021). "Multilingual Text Classification from Twitter during Emergencies". In: *2021 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–6.

Purohit, H., Castillo, C., Imran, M., and Pandey, R. (2018). "Social-EOC: Serviceability Model to Rank Social Media Requests for Emergency Operation Centers". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 119–126.

Ratner, A., Bach, S., Varma, P., and Ré, C. (2019). "Weak supervision: the new programming paradigm for machine learning". In: *Hazy Research. Available via https://dawn. cs. stanford. edu//2017/07/16/weak-supervision/. Accessed*, pp. 05–09.

Reimers, N. and Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv: 1908.10084 [cs.CL].

Rekatsinas, T., Chu, X., Ilyas, I. F., and Ré, C. (2017). "Holoclean: Holistic data repairs with probabilistic inference". In: *arXiv preprint arXiv:1702.00820*.

Reuter, C., Backfried, G., Kaufhold, M.-A., and Spahr, F. (2018). "ISCRAM turns 15: A Trend Analysis of Social Media Papers 2004-2017". In: *Proceedings of the 15th International ISCRAM Conference* May, pp. 1–14.

Reuter, C. and Kaufhold, M. A. (2018). "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis Informatics". In: *Journal of Contingencies and Crisis Management* 26.1, pp. 41–57.

Rossi, C., Acerbo, F., Ylinen, K., Juga, I., Nurmi, P., Bosca, A., Tarasconi, F., Cristoforetti, M., and Alikadic, A. (Mar. 2018). "Early Detection and Information Extraction for Weather-induced Floods using Social Media Streams". In: *International Journal of Disaster Risk Reduction* 30.

Shklovski, I., Palen, L., and Sutton, J. (2008). "Finding Community through Information and Communication Technology in Disaster Response". In: *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. CSCW '08. San Diego, CA, USA: Association for Computing Machinery, pp. 127–136.

St. Denis, L. A., Hughes, A. L., and Palen, L. (2012). "Trial by fire: The deployment of trusted digital volunteers in the 2011 shadow lake fire". In: *ISCRAM 2012 Conference Proceedings - 9th International Conference on Information Systems for Crisis Response and Management* April, pp. 1–10.

Starbird, K. and Palen, L. (2013). "Working and Sustaining the Virtual "Disaster Desk"". In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. CSCW '13. San Antonio, Texas, USA: Association for Computing Machinery, pp. 491–502.

Sun, Q., Liu, Y., Chua, T.-S., and Schiele, B. (June 2019). "Meta-Transfer Learning for Few-Shot Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (June 2020). "Generalizing from a Few Examples: A Survey on Few-Shot Learning". In: *ACM Comput. Surv.* 53.3.

White, J. I., Palen, L., and Anderson, K. M. (2014). "Digital Mobilization in Disaster Response: The Work and Self-Organization of on-Line Pet Advocates in Response to Hurricane Sandy". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '14. Baltimore, Maryland, USA: Association for Computing Machinery, pp. 866–876.

Xia, R., Zong, C., Hu, X., and Cambria, E. (2015). "Feature Ensemble plus Sample Selection: Domain Adaptation for Sentiment Classification". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, pp. 4229–4233.

Zhang, C., Ré, C., Cafarella, M., De Sa, C., Ratner, A., Shin, J., Wang, F., and Wu, S. (2017). "DeepDive: Declarative knowledge base construction". In: *Communications of the ACM* 60.5, pp. 93–102.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). "Learning with local and global consistency". In: *Advances in Neural Information Processing Systems 16*. MIT Press, pp. 321–328.