Social-EOC: Serviceability Model to Rank Social Media Requests for Emergency Operation Centers

Hemant Purohit
George Mason University
Fairfax, VA, USA
hpurohit@gmu.edu

Carlos Castillo
Universitat Pompeu Fabra
Barcelona, Spain
chato@acm.org

Muhammad Imran
Qatar Computing Research Institute
Doha, Qatar
mimran@hbku.edu.qa

Rahul Pandey
George Mason University
Fairfax, VA, USA
rpandey4@gmu.edu

***Abstract*—The public expects a prompt response from emergency services to address requests for help posted on social media. However, the information overload of social media experienced by these organizations, coupled with their limited human resources, challenges them to timely identify and prioritize critical requests. This is particularly acute in crisis situations where any delay may have a severe impact on the effectiveness of the response. While social media has been extensively studied during crises, there is limited work on formally characterizing serviceable help requests and automatically prioritizing them for a timely response.**

In this paper, we present a formal model of serviceability called *Social-EOC* (Social Emergency Operations Center), which describes the elements of a *serviceable* message posted in social media that can be expressed as a request. We also describe a system for the discovery and ranking of highly serviceable requests, based on the proposed serviceability model. We validate the model for emergency services, by performing an evaluation based on real-world data from six crises, with ground truth provided by emergency management practitioners. Our experiments demonstrate that features based on the serviceability model improve the performance of discovering and ranking ($nDCG$ up to 25%) service requests over different baselines. In the light of these experiments, the application of the serviceability model could reduce the cognitive load on emergency operation center personnel, in filtering and ranking public requests at scale.

***Keywords*-Information Overload, Serviceability, Social Media, Emergency Management, Help Intent**

I. Introduction

Social media plays a key role in connecting public to all kinds of organizations, including governments, nonprofits, and for-profits. In for-profit companies, recent years have demonstrated the value of extending their customer relationship services to social media [1]. They often provide timely answers to social media queries from existing and potential customers. Similarly, recent studies show that the public expects a timely response to queries on social media addressed to governments and nonprofits [2]–[4].

From the perspective of emergency services, however, there are substantial challenges for meeting these expectations. There are vast amounts of messages posted with high velocity in social media by the public during emergencies, leading to information overload in emergency services [5], [6], given their limited human resources. Messages are also extremely varied in their potential value for operational response, ranging from actionable requests or concrete offers of help [7]–[9] to unsubstantiated rumors [10]. Thus, quickly prioritizing messages with help seeking intent that require a timely response has become a critical need for agencies in the emergency operation centers (EOCs) [11].

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00

TABLE I
Example messages with different serviceability characteristics addressed to Fort Bend County Office of Emergency Management (@fbcoem) in the US during Hurricane Harvey in 2017. *(Messages rephrased for anonymity)*

	Message	Characteristics
M1	@fbcoem I am 9 ft above current water levels, why am I told to evacuate Grand Lakes now? Please advise.	serviceable
M2	@fbcoem If there has been no rain since yesterday, why is water not draining?	serviceable, lacks details
M3	@fbcoem Thank God you are working on this. Let us chat when things settle down	not serviceable

Table I shows some example messages addressed to Fort Bend County Office of Emergency Management in the US. *M1* is a prototypical serviceable message, containing a concrete request (confirmation of evacuation order). *M2* is still serviceable, it has a request for information, however there is ambiguity (where is the water not draining?) that makes it less serviceable. Finally, *M3* is not a serviceable message for operational response, but a message expressing gratitude.

Problem. We address the problem of filtering and prioritizing serviceable social media requests for emergency services.

Our contribution. To our knowledge, this is a first study to formally define a generic *serviceability* model (*Social-EOC*) for social media messages, in order to identify and prioritize actionable requests to respond for emergency services. We introduce the model in Section III together with a qualitative and quantitative description of serviceability characteristics. A learning-to-rank system based on the proposed model is implemented in Section IV, using inferences of the serviceability characteristics. This system classifies and ranks serviceable requests, for ultimately reducing the cognitive load on EOC personnel in processing public requests at large scale.

Finally, we demonstrate the validity of the *Social-EOC* model in Sections V and VI, by experimenting with the real-world datasets from six crises during the past six years. We present our conclusions in Section VII.

II. Related Work

Literature on the topic of social media for emergency management is vast, for a survey, see [6], [12]. In this section we outline closely related work, focusing on research done for determining what is serviceable (or even relevant and actionable) for emergency services.

A. Social media during emergencies

"Big Crisis Data" from social media has such a high volume, variety, and velocity, that it can overwhelm response services [12]. Crisis informatics [13] has investigated the use of social media for emergency services. Quantitative approaches have focused on studying public behavior in specific emergencies while addressing problems of data collection and filtering, classification, summarization as well as visualization [6]. Prior research has identified information overload as a key challenge and a barrier for the efficient use of social media communication by emergency services [12], [14]. Information overload originates from a variety of factors including the large scale, unstructured, and noisy nature of social media content. Furthermore, the characteristics of relevant social media requests that must be prioritized are not well understood.

B. Services in emergency management

In the emergency management domain, Public Information Officers (PIOs) play the role of serving information to the public or sourcing relevant information from public sources for the response agencies or an EOC, by leveraging various information communication technologies including social media [15]. PIOs are provided guidelines for communication with the public [16], and have the responsibility to communicate critical information and respond promptly to requests. Over the last few years, PIOs have increasingly used social media to communicate effectively with the public. Reports and surveys of emergency services [4], [11], [17] recognize social media as a valuable information channel for improving their operational response coordination, however, they also recognize the necessity to effectively filter, prioritize, and organize information from this channel.

C. Mining intent of requesting help

The literature provides some guidance on modeling requesting behavior or information seeking intent across different domains [18], including Q&A forum [19], email communication [20], and social media platforms [7], [21], [22]. In online fora, researchers have found request behavior in varied contexts such as urgency, informational intent, and social support. However, prior research on information seeking behavior is generic for all types of users and often not targeted towards seeking answers from a specific agency, organization, or group of organizations, as we focus in this study. In email communications [20], researchers explored the characteristics of ranking messages for replying and created predictive models for prioritization. However, the length of emails provide a greater context to express the request behavior, which does not apply to typically shorter social media messages. The most relevant line of work for our analysis is request behavior on social media, which has been defined by researchers as ranging from explicit requests for organizational users [22] to implicit requests for seeking donations and resources [7], [8], [23] and other actions [9] during disasters. In particular for explicit requests to agencies, Sachdeva et al. [22] defined requests to which police agencies should respond, evaluate, or take action as serviceable requests, by analyzing the messages on a Facebook page of a police department. Ferrario et al. [21] analyzed the #bbcqt hashtag used for BBC Question Time (a current affairs discussion program broadcast on BBC One in the UK) to find actionable tweets. References [7], [8], [24] proposed methods to identify implicit request messages for seeking or offering to help resources during disaster relief, however, not specifically targeted to emergency services. Ranganath et al. [9] created a method to identify users who can provide timely and relevant responses to actionable questions posted on social media, but not specific factors for organizational agency users. To complement the prior research on social media for request behavior, we focus on creating a generalizable model for serviceability characteristics of requests targeted to organizational services.

III. *Social-EOC*: Serviceability Model

In this section we describe a qualitative model of serviceability, followed by a quantitative model.

A. Qualitative serviceability model

We consider a general class of emergency service requests, following official guidelines from the US FEMA (Federal Emergency Management Agency) [16], which include intended *actions* such as a request for resources (e.g., emergency medical assistance for an injured person) as well as *information* (e.g., a request for a phone number to get information on missing people). The key characteristic of a serviceable request message is that it requests a resource that can be provided, or asks a question that can be answered. For instance, we do not consider messages whose sole purpose is to congratulate/praise or complaint as serviceable; in our framework, a serviceable message must contain an explicit request for resources or a concrete question.

The serviceability of a social media message is also determined by whether it is correctly addressed to an organization that can provide the resource or information. Most social media platforms include features for sending publicly or privately a message addressed to a specific user. Thus, a citizen seeking an action or answer from an organization can address the request to that organization's account.

We note that each organization or agency usually has its own protocols to determine if and how a request should be answered. However, the knowledge of such protocols is acquired by the service personnel through training guidelines, and may remain in the form of tacit knowledge instead of structured knowledge that could be used to automate responses.

Finally, serviceability not only refers to the topicality of the request and to the fact that it must be addressed correctly,

TABLE II
SUMMARY OF DATASETS FOR TWEETS CONTAINED IN THE CONVERSATIONS, THE SAMPLED TWEETS FOR ANNOTATION, AND THE DISTRIBUTION OF EXPERT LABELS FOR SERVICEABLE REQUESTS. "TARGETS" REFER TO OUR LIST OF ACCOUNTS OF GOVERNMENT AND NONPROFIT EMERGENCY-RELATED ORGANIZATIONS. NOTICE THAT FOR THE FIRST THREE EVENTS WE HAVE LESS THAN 100 LABELED DATA POINTS, WHILE FOR THE NEXT THREE EVENTS WE HAVE SEVERAL HUNDRED LABELS PER EVENT.

Event (start-end month/day)	Conversational Tweets	Sampled Tweets containing targets	Serviceable	Not Serviceable
Hurricane Sandy 2012 (10/27-11/07)	1,153	60	24 (40%)	36 (60%)
Oklahoma Tornado 2013 (11/07-11/17)	1,513	52	25 (48%)	27 (52%)
Louisiana Floods 2016 (10/11-10/31)	1,369	56	19 (34%)	37 (66%)
Alberta Floods 2013 (06/16-06/16)	2,727	814	229 (28%)	585 (72%)
Nepal Earthquake 2015 (04/15-05/15)	2,222	240	43 (18%)	197 (82%)
Hurricane Harvey 2017 (08/29-09/15)	12,742	1,534	306 (20%)	1,228 (80%)

but also to whether it contains required details, such as time, place, or context. In summary, we propose the following definition of a request on social media with the serviceability characteristics.

Definition 1: **serviceable request**. A *serviceable request* in social media is a message that: (i) requests a resource that can be provided or asks a question that can be answered, (ii) addresses a person or organization that can provide the resource/answer, and (iii) provides sufficient details for the resource/answer to be provided.

B. Quantitative serviceability model

Our definition 1 describes an ideal serviceable message, but serviceability is a matter of degree. To quantify this, we associate a score to each of the three types of serviceability characteristics for a given message m, for instance by using a 5-points Likert Scale [25]:

Explicit request/answerable question. Two scores:

- a score ($E(m)$) for the characteristic of being an *Explicit Request*, i.e., ideally a message that explicitly asks for a resource or service, e.g., message *M1* in Table I.
- a score ($A(m)$) for the characteristic of being an *Answerable Question*, i.e., ideally a request message that explicitly asks a question that can be answered, e.g., messages *M1* and *M2* in Table I.

Correctly addressed. a score ($C(m)$) for the characteristic of being *Correctly Addressed*, i.e., ideally a message sent to (addressed or mentioning) the person or organization who could have the resource, or could provide the service, an alarm, or could answer the question, e.g., messages *M1, M2*, and *M3* in Table I.

Sufficiently detailed. a score ($D(m)$) for the characteristic of providing *Sufficiently Detailed* context, i.e., ideally a message specifying enough contextual information such as time (when), location (where), quantity (how much), sub-type of resource (which), to make the request or question unambiguous, e.g., message *M1* in Table I.

Our quantitative serviceability model for a message is defined as a function of these characteristics $f(E(m), A(m), C(m), D(m))$. We describe its implementation next.

IV. IMPLEMENTING THE SERVICEABILITY MODEL

The proposed system, implementing the *Social-EOC* model, depends primarily on four steps:

A) Collecting conversation streams
B) Rating serviceability characteristics
C) Creating gold standard of serviceable requests
D) Learning to classify and rank serviceable requests

We present details of each of them in the following.

A. Collecting Conversation Streams

We first collected data from Twitter for six disaster events from the last six years, using the keyword-based crawling approach. We collected tweets during hurricane Harvey in 2017 and Louisiana floods in 2016 using CitizenHelper system [26] and for prior events, re-used datasets available from previous works [27], [28]. Following the recommendation of collecting "contextual streams" [29], we further extended each event collection with messages that belonged to conversation chain (a *Reply* message thread on Twitter), where a conversation chain contained at least one message from an event dataset. To collect such conversation chain messages, we "scrapped" web pages of conversations using *tweet id* in each of our event datasets (this allows to recover more public messages than using Twitter's API, which does not provide conversation chains). Specifically, the conversation chain for tweet with id *TWEETID* authored by a user with handle *USER* is available at URL *http://twitter.com/USER/status/TWEETID*. Table II shows a summary of the final dataset.

B. Rating Serviceability Characteristics

We asked crowdsourcing annotators for rating the individual serviceability characteristics of a message, as we describe in this section. We also requested domain practitioners for the ground truth annotation of the overall serviceability of a message (described in the next section.)

For rating individual serviceability characteristics, we provided instructions and examples to the crowdsourcing annotators (specifically, university student volunteers) based on the model described in Section III. Given a message, three annotators associated a numerical rating between 1 to 5 to each serviceability characteristic. We also solicited the rating on an additional attribute of the message to indicate non-serviceable aspects such as complaints, gratitude,

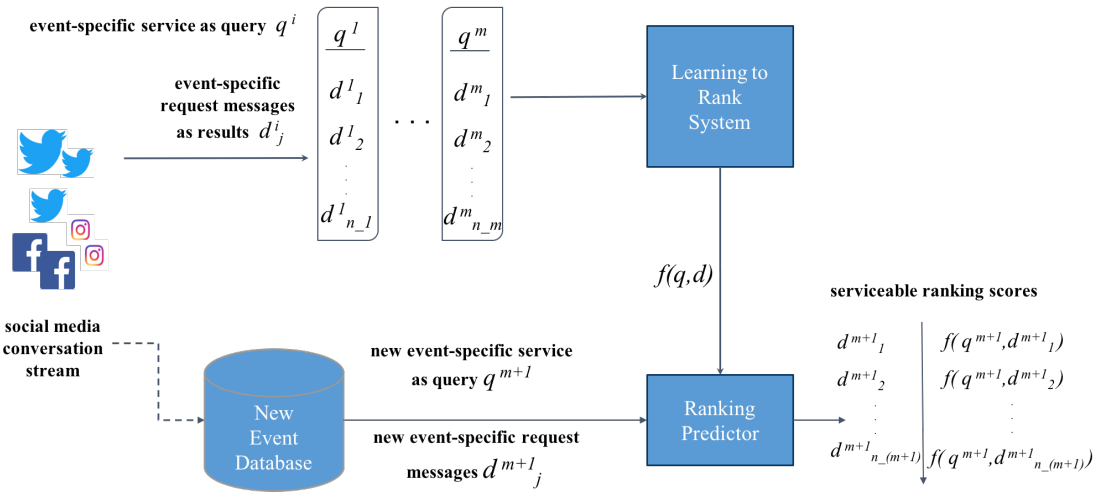

Fig. 1. Depiction of the overall system design based on *Social-EOC* model and using the learning-to-rank approach for prioritization.

congratulations, and advertisements because these are not a priority for operational response. For example, through this message "*@account1 wow that photo made me tear up! Just amazing what u do! #yycflood*", a user is only trying to praise and express gratitude towards the officials. It is not asking for any operational resource or requesting any information. We provided the following annotation task description:

Question. What are the characteristics of the information in the following message?
Instruction. Please choose the appropriate rating between 1 to 5, with 5 being the highest value.
Message. {content of message}

- **Explicit request** for a resource or service: 1 . . . 5
- **Answerable** question that could be responded: 1 . . . 5
- **Correctly addressed** to an organization or person capable of servicing or providing resources: 1 . . . 5
- **Sufficiently detailed** with contextual information for time, location, and incident: 1 . . . 5
- **Other** such as complaint, gratitude, congratulatory praise, sarcasm, and advertisement: 1 . . . 5

Examples. Multiple examples with reasoning for the possible ratings are provided, e.g., "*@account when you say appropriate footwear, does that mean good walking shoes or rubber boots? Prepared to walk through mud?*". It is requesting information with a specific intent to clarify details regarding a prior announced message of the agency (@account). Thus, it receives the highest ratings of 5 with respect to being answerable, correctly addressed, and providing sufficient details.

We also indicated that for serviceability, the requested resource or service action should be external to the social media platform, i.e., it excludes actions that are done only within the platform itself, such as "RT me", "follow me" or "read this" or "check out." An example of such a message is "*@account No matter where in the world your followers live, u can donate from here: _url_ Help #Nepal! Pls RT!* ".

For the annotation task on the conversational dataset of an event, we selected a biased set of messages using the following two equal sized samples, in order to increase the recall of potential serviceable requests. The first sample selects all the messages that were directly addressed or targeted to official accounts (i.e., that start with '*@account*') and that were posted in a conversation chain before an official reply was posted. The second sample randomly selects messages from the remaining event dataset after excluding the first sample and the messages authored by official accounts. We collected the set of official accounts of relevant response organizations for an event through the official reports, including those from FEMA and news sources. For example, *@account* in the examples of Table I is the Twitter user account of the emergency management unit of a county government responding to hurricane Harvey (the target official accounts of each event are provided in our data release).

After execution of the event-specific annotation task by three annotators per message, we computed the average ratings of characteristics per message. Table III shows examples of messages with average ratings.

C. Creating Gold Standard of Serviceable Requests

To validate our model for serviceability, we required a gold standard set of serviceable requests. The set was designed with the help of domain experts by labeling the annotated message set from the previous section. For this task, we asked domain experts to label if a request message would qualify as *serviceable* or *not serviceable*, according to their experience. We also provided them an optional field for entering comments on their choice of label. Our domain experts are three active professionals in the emergency management domain located in the United States, Canada, and Nepal, who have had roles in the public communications in a response agency. Specifically, the US-based expert labeled messages from Harvey, Louisiana, Oklahoma, and Sandy events, the expert based in Canada

TABLE III
EXAMPLE MESSAGES WITH THE RATINGS GIVEN BY ANNOTATORS, CONSIDERING THE SERVICEABILITY CHARACTERISTICS OF MESSAGES. (MESSAGE TEXT PARAPHRASED FOR ANONYMITY.)

Message	Explicit Request	Answerable Question	Correctly Addressed	Sufficiently Detailed
M4: @account1 plz, governor, post a phone # for specific info in our local areas	4.33	4.33	3.33	3.67
M5: @account2 is thr parking at McMahon for volunteer?	4.00	5.00	5.00	5.00
M6: @account3 how can I help	1.30	4.33	4.33	1.00
M7: @account4 Plz pray for these families	1.66	1.00	1.00	1.00
M8: @account5 been working in #LAFlood @account6 shelter, we actively monitor Social Media for feedback	1.00	1.00	2.00	2.00

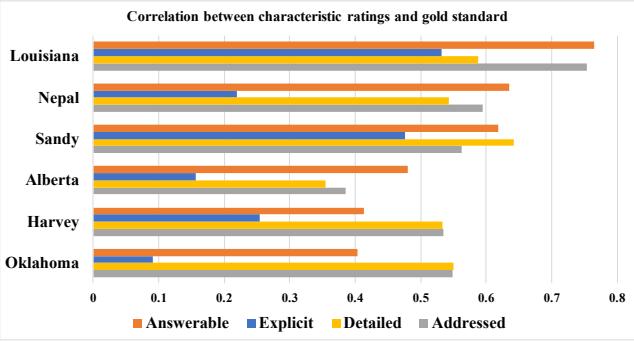

Fig. 2. We observe a positive correlation between the average crowdsourced (non-expert) ratings of the proposed serviceability characteristics and the gold standard annotations of serviceability given by domain experts, according to Pearson Correlation. All correlations are statistically significant at $p \leq 0.01$ level (2-tailed) except for the case of Explicit Request in the Oklahoma event.

labeled the Alberta event, and the expert based in Nepal labeled the Nepal earthquake event.

Table II summarizes the resultant label distribution from the three domain experts. The comments they provided during their label annotation process show some insights.

First, experts in general considered that serviceable messages were a subclass of the messages they would respond to during a disaster; in other words, that they would not answer only to requests for actions or information, but sometimes, also to other classes of messages, for example *M8* in Table III.

Second, a disagreement between the experts was observed in relation to messages that only express gratitude (such as M3 in Table I). One of the three experts considered such messages as serviceable, with the rationale that replying to gratitude messages could help strengthen trust within a community. In contrast, the other two experts considered gratitude messages as not serviceable, as they did not consider them a priority for operational response like the other actionable messages. We sided with the majority opinion and resolved to keep the gratitude messages under the *not serviceable* category.

Third, experts identified that in some cases a message should be answered to provide reassurance or restate facts. This was considered a good strategy for countering rumors, in particular highly alarming or easily falsifiable ones.

Overall, we found a correlation between the gold standard annotations of serviceability by experts and the average ratings by non-expert crowdsourcing workers for the proposed serviceability characteristics, as shown in Figure 2.

D. Learning to Classify and Rank Serviceable Requests

For filtering and prioritizing serviceable requests, we propose a supervised learning method for automatic classification and ranking. In automatic classification, the objective is to classify a message as either *serviceable* or *not serviceable*. In automatic ranking, the objective is to order a list of messages according to how serviceable they are. The *learning-to-rank* methodology [30] is suitable for jointly meeting these objectives. This method could use any kind of relevance levels of serviceability for training purpose. In our experiments, we have considered only binary levels, but we remark the method is general. Figure 1 depicts how learning-to-rank fits within this process, as we explain next.

Formally, we consider that each event $i = 1, 2, \ldots, m$ determines a *query* q^i, and $D^i = \{d^i_1, d^i_2, \ldots, d^i_{n\_i}\}$ are all the messages relevant to that event, which need to be ranked. Each message is associated to a label in $Y = \{y_1, y_2, \ldots, y_\ell\}$ representing its level of serviceability (a total order between the graded levels exist). The training set corresponds to tuples $\langle q^i, D^i, Y^i \rangle$ containing queries, documents for each query, and sets of labels for each document, with $(Y^i)_j$ indicating the label for document d^i_j.

In this context, the goal of a learning-to-rank method is to learn a ranking model that associates to each query q^i a permutation of the documents in D^i that matches as much as possible the training labels, in the sense that higher graded labels in Y are associated to documents appearing near the top of the ranking (being more *serviceable*). In particular, we consider learning-to-rank models characterized by functions $f_i : D^i \times D^i \to \{-1, +1\}$ that associate to each pair of documents $(u, v) \in D^i$ a score -1 if u should be ranked above v for q^i and $+1$ otherwise. Specifically, $f_i(u, v) = -1$ if $(Y^i)_u > (Y^i)_v$, $f_i(u, v) = +1$ if $(Y^i)_v > (Y^i)_u$. To solve this problem, we use the SVM-Rank algorithm [31].

Feature Extraction. Query-document pairs (q^i, d^i_j) are represented by feature vectors, which include the following:

- *Generic features*: counts of the number of words, hashtags, user mentions, and URLs in a tweet.
- *Text features*: *tf-idf* for a bag-of-words, after performing standard text-preprocessing on a request message (removing non-ASCII characters, tokenization, removing stopwords, removing URLs) and lastly, replacing number, retweet indicator (`RT @USER`), and mention indicator (`@USER`) with tokens `_num_`, `_rt_`, `_mention_`.

- *Serviceability features*: we consider two sources of features for serviceability characteristics.
  - manual labels: numerical scores between 1 to 5 for the average rating of each serviceable characteristic (explicit, answerable, etc.) provided by crowdsourcing workers in the annotation task.
  - inferred labels: binary scores generated by an automatic classifier for each characteristic (explicit, answerable, etc.). The automatic classifier was created using logistic regression, with features that are pre-trained `word2vec` representations of the messages (embedding size 300) taken from *Google Word2Vec toolkit*, which is trained on continuous bag-of-words architecture [32]. Training includes a held-out portion of messages for the classes of 0 and 1, corresponding to the manual labels $\{1, 2\}$ and $\{3, 4, 5\}$ respectively.

V. Experimental Setup and Evaluation

For a robust validation of the proposed serviceability model of request messages, we compare the following classification and ranking schemes for requests. In all cases, generic features are computed and text is pre-processed as described in Section IV-D.

- **[*T*]: Text + generic features** (baseline). This method uses a bag-of-words (BoW) representation of the text features, along with the generic features.
- **[*T+I*]: T + inferred labels**. This method uses features from *T* plus serviceability characteristic labels generated by an automatic classifier trained on a held-out portion of messages from each event.
- **[*T + I_all*]: T + inferred from all-events model**. This method uses features from *T* plus serviceability characteristic labels generated by automatic classifiers trained on a held-out portion of messages from all 6 events.
- **[*T + I_cross*]: T + inferred from cross-event model**. This method is similar to *T + I_all* but only 5 events are used for training the automatic classifier of serviceability characteristic labels (the held-out portion of messages for the event being considered in each case is not included).
- **[*T_cross + I_cross*]: T cross-events + inferred from cross-event model**. This method computes the model for *T* and for inferred serviceability characteristics using 5 events (excluding the event being considered).
- **[*T+M*]: T + manual labels** (hand-labeled). This method uses features from *T* plus serviceability characteristic labels provided by crowdsourcing annotators. It represents a "best-case" scenario in which each message already has been annotated manually along each serviceability characteristic, which is not realistic in a real-world situation with a large-volume dataset.

Evaluation metrics. To compare the different methods we use a popular measure from Information Retrieval: the normalized Discounted Cumulative Gain ($nDCG$) [30], which effectively compares two rankings by weighing more differences in the top positions than differences further down. Specifically, for each event/query:

$$nDCG(k) = G^{-1}_{max,i}(k) \sum_{j:\pi_i(j)\leq k} \frac{2^{y_{i,j}} - 1}{\log_2(1 + \pi_i(j))}$$

where

- $\pi_i(j)$: Position of the document d^i_j in ranking list π_i
- $G^{-1}_{max,i}(k)$: Normalizing factor at position k
- $y_{i,j}$: label of document d^i_j in ranking list π_i

In this study, we analyzed $nDCG$ for the top-5 and top-10 ranked messages, for the rankings obtained across each fold of the 5-fold cross validation setting, for each event.

VI. Results and Discussion

In this section we present the results from the proposed ranking schemes, comparing their performance with respect to the features of the serviceability model. We then discuss the limitations of this approach and future work directions.

TABLE IV
Comparison of $nDCG@5$ and $nDCG@10$ (expressed as percentages) using 5-fold cross validation for each event. For the small datasets, we have less than one hundred tweets with serviceability (expert-provided) labels, while we have several hundred labels for each event in the large datasets. Bold values indicate the best performing scheme among the ones using inferred serviceability characteristics.

Event	Classification Schemes	$nDCG@5$	$nDCG@10$
	Small datasets ($n_{\texttt{labeled}} < 100$)		
Oklahoma	T *(baseline)*	49%	74%
	T+I	**53%**	**77%**
	T+I_all	46%	72%
	T+I_cross	42%	71%
	T_cross+I_cross	46%	72%
	T+M *(hand-labeled)*	*61%*	*85%*
Sandy	T *(baseline)*	50%	67%
	T+I	57%	75%
	T+I_all	57%	75%
	T+I_cross	**71%**	**87%**
	T_cross+I_cross	56%	79%
	T+M *(hand-labeled)*	*72%*	*90%*
Louisiana	T (baseline)	94%	96%
	T+I	89%	96%
	T+I_all	89%	96%
	T+I_cross	**96%**	**99%**
	T_cross+I_cross	77%	90%
	T+M *(hand-labeled)*	*93%*	*98%*
	Large datasets ($n_{\texttt{labeled}} \geq 100$)		
Nepal	T (baseline)	46%	44%
	T+I	52%	50%
	T+I_all	55%	50%
	T+I_cross	52%	50%
	T_cross+I_cross	**58%**	**63%**
	T+M *(hand-labeled)*	*74%*	*66%*
Harvey	T (baseline)	62%	60%
	T+I	64%	62%
	T+I_all	62%	**69%**
	T+I_cross	**64%**	62%
	T_cross+I_cross	54%	52%
	T+M *(hand-labeled)*	*74%*	*78%*
Alberta	T (baseline)	57%	47%
	T+I	56%	52%
	T+I_all	49%	53%
	T+I_cross	56%	52%
	T_cross+I_cross	**65%**	**59%**
	T+M *(hand-labeled)*	*91%*	*84%*

A. Result observations

Table IV compares the performance of the different methods in terms of $nDCG$ of the first 5 positions ($nDCG@5$) and 10 positions ($nDCG@10$). We observe the following:

- **The serviceability characteristics of our model capture the notion of serviceability.** The performance of the methods based on inferring serviceability characteristics is above the performance of the baseline, and if serviceability characteristics are given as inputs (method ***T+M***, hand-labeled), we obtain the best performance. We note, however, that obtaining labels for serviceability characteristics from human annotators in real-time is not practical; hence, we need to use inferred characteristics.
- **Inferring serviceability characteristics is better than the baseline.** Overall, there is a consistent pattern of good performance across the proposed ranking systems (***T+I*** and variants). The improvement in $nDCG@5$ and $nDCG@10$ is obtained by adding the proposed serviceable characteristics features (automatically inferred) to the baseline text (bag-of-words) features (***T***).
- **Performance varies in cases of small datasets.** In the small datasets, we note that in the Louisiana event, for top-5, (***T+I***) results are not better than ***T***. Note that given we use 5-fold cross-validation, in the case of small datasets we have at most $60/5 = 12$ examples per fold, and in the case of large datasets we have $240/5 = 48$ examples per fold at least. We also note there are less training examples, which is in line with our next observations.
- **Cross-event models perform well.** We found promising results for approaches that use cross-event datasets to create a model for serviceability characteristics. In particular, the performance gains are clearer in the case of smaller event datasets than the larger ones, given the possibility of learning from a larger corpus. This is evident from the results for the Harvey event, ***(T_cross + I_cross)***, given that in this dataset we have more labeled data than all other events combined.
- **Serviceability characteristics based features are among the best discriminators**. In the ***T+I*** model and variants, the inferred serviceability features were consistently among the top-5 features of the classifiers. We identified the top-5 features using χ^2 test with stratified 5-fold cross validation. This feature analysis further justifies the improvement over the baseline model in Table IV.

Table V shows the top ("best") and bottom ("worst") ranked messages using the ***T + I*** method based on text (bag-of-word) and inferred serviceability characteristics. The examples match the expectations on qualitative serviceability characteristics.

TABLE V
TOP-2 (MOST SERVICEABLE) AND BOTTOM-2 (LEAST SERVICEABLE) TWEETS FOR EACH EVENT, OBTAINED AUTOMATICALLY USING THE ***T+I*** EXPERIMENTAL SCHEME.

Event	Ranked Requests
Sandy	
[TOP]	- please, governor, post a website or phone# where we can get specific info for our local areas
	- Queens trains aren't being addressed at all. When can we expect any service updates for the NQR trains?
[BOTTOM]	- Romney not going 2 like that Gov Christie is being nice about Obama's leadership #sandy
	- HILARIOUS! That is much needed laughter, I am sure.
Oklahoma	
[TOP]	- how can I donate to the US red cross from the UK? No option to donate from a UK address on the site
	- you are correct only a perc goes to the victims
[BOTTOM]	- Sending up many prayers from here in Colorado for everyone
	- help us too!!! #oklahoma
Alberta	
[TOP]	- can you tell me if sanitary pumps are running yet in elbow park? #yycflood
	- plz text with what you need & address. Lots of volunteers in mission
[BOTTOM]	- thank u calgary police
	- Tx for ur time!!
Nepal	
[TOP]	- able to organise a collection of goods you mention but can u guarantee the capacity on ground
	- cc @account0 have a collection point here in Hyd. Let me see if I can tag u to another tweet.
[BOTTOM]	- plz send 100 bits for #NepalEarthquake disaster recovery
	- I'm trying to send a fiver, FFS...
Louisiana	
[TOP]	- help needed! #Laflood #children #teens #teachers URL
	- Want to help those who were affected by #LouisianaFlood, join us #LouisianaStrong URL
[BOTTOM]	- Obama has taken fewer vacation days than any in recent. Wud u prefer he work 24/7
	- why didn't Jesus prevent the flood pretty simple no ?
Harvey	
[TOP]	- This list of neighborhoods has caused confusion, we need clarity on exactly what areas are impacted by this.
	- What percent of donation goes directly to aid? it affects where I give at this time.
[BOTTOM]	- yes, thank you #harvey
	- I did thank you!

B. Limitations and Future Work

While the applied method can prioritize serviceable messages, it has some limitations.

First, the analyzed data contains only English language messages, which is the dominant language in the collections we use. However, we anticipate the core information characteristics of serviceability for the messages to be the same, or similar, across messages written in other languages. Similarly, we note that all our datasets come from the same platform, i.e. Twitter, and we would need to evaluate messages in other platforms where populations, functionalities, and norms may be different and hence, serviceability might differ.

Second, we only limited our experiments to conversational messages starting with or a mention of a user account (i.e., addressing an account) that might be asking a potential request to an agency or organization. This was done to reduce the amount of ground-truth labeling required, given the limited time of domain experts. It is possible that users may indirectly ask the agencies for a service request, without mentioning their user account explicitly. Therefore, future work could expand the presented analytical results with such indirect requests.

Third, our experiments considered overall serviceability at

the binary levels. Future works could consider serviceability to be a matter of non-binary grading levels; however, the same methodology and evaluation could be applied given that $nDCG$ can be used with non-binary relevance assessments.

VII. Conclusions

This paper presents a novel study of serviceable request characteristics of social media messages, which can improve prioritization and filtering of messages from social streams for emergency services. This is done through a novel model for serviceability of social media requests sent to an organization or agency, called the *Social-EOC* serviceability model.

We demonstrated the applicability of this model for emergency services by creating different types of classification and ranking systems using the proposed serviceability characteristics of a request message. Specifically, we proposed several systems for classifying and ranking requests for serviceability, with a baseline text-based method using bag-of-word features, and a series of variants of our method using inferred features for the serviceability characteristics.

Our experimental evaluation on six disaster events showed a consistent performance gain for the systems that were based on inclusion of features for the serviceability characteristics (relative gain in $nDCG@10$ and $nDCG@5$ of up to 25%). The application of the proposed method can help in improving social media services at emergency management organizations. This in turn can provide a complementary capability for traditional communication channels such as 911 in the United States and 112 in Europe that get often overwhelmed during mass emergencies.

Reproducibility. Anonymized messages, a list of official accounts, and the crowdsourced and expert labels are available at http://ist.gmu.edu/~hpurohit/informatics-lab/crisis-data.html

VIII. Acknowledgement

Purohit thanks US National Science Foundation grant IIS-1657379 and Castillo thanks La Caixa project LCF/PR/PR16/11110009 for partial support. Authors would also like to acknowledge reviewers for valuable feedback.

References

[1] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? get serious! understanding the functional building blocks of social media," *Business horizons*, vol. 54, no. 3, pp. 241–251, 2011.

[2] American Red Cross, "More americans using mobile apps in emergencies," 2012, online and phone survey. [Online]. Available: http://www.redcross.org/news/press-release/More-Americans-Using-Mobile-Apps-in-Emergencies

[3] A. L. Hughes, L. A. St Denis, L. Palen, and K. M. Anderson, "Online public communications by police & fire services during the 2012 hurricane Sandy," in *Proc. ACM SIGCHI*, 2014, pp. 1505–1514.

[4] C. Reuter and T. Spielhofer, "Towards social resilience: A quantitative and qualitative survey on citizens' perception of social media in emergencies in europe," *Technological Forecasting and Social Change*, vol. 121, pp. 168–180, 2017.

[5] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, "Using social media to enhance emergency situation awareness," *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 52–59, 2012.

[6] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.

[7] H. Purohit, C. Castillo, F. Diaz, A. Sheth, and P. Meier, "Emergency-relief coordination on social media: Automatically matching resource requests and offers," *First Monday*, vol. 19, no. 1, 2013.

[8] X. He, D. Lu, D. Margolin, M. Wang, S. E. Idrissi, and Y.-R. Lin, "The signals and noise: Actionable information in improvised social media channels during a disaster," in *Proc. ACM WebSci*, 2017, pp. 33–42.

[9] S. Ranganath, S. Wang, X. Hu, J. Tang, and H. Liu, "Facilitating time critical information seeking in social media," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2197–2209, 2017.

[10] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason, "Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing," *Proc. iConference*, 2014.

[11] U.S. Homeland Security, "Using social media for enhanced situational awareness and decision support," 2014. [Online]. Available: https://www.dhs.gov/publication/using-social-media-enhanced-situational-awareness-decision-support

[12] C. Castillo, *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press, 2016.

[13] L. Palen and K. M. Anderson, "Crisis informatics – new data for extraordinary times," *Science*, vol. 353, no. 6296, pp. 224–225, 2016.

[14] S. R. Hiltz, J. A. Kushma, and L. Plotnick, "Use of social media by us public sector emergency managers: Barriers and wish lists." in *Proc. ISCRAM*, 2014, pp. 602–611.

[15] A. L. Hughes and L. Palen, "The evolving role of the public information officer: An examination of social media in emergency management," *Journal of Homeland Security and Emergency Management*, vol. 9, no. 1, 2012.

[16] FEMA, "Public information officer (PIO)," https://training.fema.gov/programs/pio/, 2017.

[17] U.S. Homeland Security, "From concept to reality: Operationalizing social media for preparedness, response and recovery," 2016. [Online]. Available: https://www.dhs.gov/publication/vsmwg-concept-reality

[18] J.-E. Mai, *Looking for information: A survey of research on information seeking, needs, and behavior*. Emerald Group Publishing, 2016.

[19] B. Vasilescu, A. Serebrenik, P. Devanbu, and V. Filkov, "How social Q&A sites are changing knowledge sharing in open source software communities," in *Proc. ACM CSCW*, 2014, pp. 342–354.

[20] L. Yang, S. T. Dumais, P. N. Benne, and A. H. Awadallah, "Characterizing and predicting enterprise email reply behavior," in *Proc. ACM SIGIR*, 2017, pp. 235–244.

[21] M. A. Ferrario, W. Simm, J. Whittle, P. Rayson, M. Terzi, and J. Binner, "Understanding actionable knowledge in social media: BBC question time and Twitter, a case study." in *Proc. ICWSM*, 2012, pp. 455–4458.

[22] N. Sachdeva and P. Kumaraguru, "Call for service: Characterizing and modeling police response to serviceable requests on Facebook." in *Proc. ACM CSCW*, 2017, pp. 336–352.

[23] I. Varga, M. Sano, K. Torisawa, C. Hashimoto, K. Ohtake, T. Kawai, J.-H. Oh, and S. De Saeger, "Aid is out there: Looking for help from tweets during a large scale disaster," in *Proc. ACL*, vol. 1, 2013, pp. 1619–1629.

[24] T. H. Nazer, F. Morstatter, H. Dani, and H. Liu, "Finding requests in social media for disaster relief," in *Proc. IEEE/ACM ASONAM*, 2016, pp. 1410–1413.

[25] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[26] P. Karuna, M. Rana, and H. Purohit, "Citizenhelper: A streaming analytics system to mine citizen and web data for humanitarian organizations." in *Proc. ICWSM*, 2017, pp. 729–730.

[27] A. Sheth, A. Jadhav, P. Kapanipathi, C. Lu, H. Purohit, G. A. Smith, and W. Wang, "Twitris: A system for collective social intelligence," in *Encyclopedia of social network analysis and mining*. Springer, 2014, pp. 2240–2253.

[28] M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: human-annotated Twitter corpora for NLP of crisis-related messages," in *Proc. LREC*, 2016.

[29] L. Palen, "Frontiers of crisis informatics," Computer Science Colloquia, University of Colorado, Boulder, 2014. [Online]. Available: https://www.cs.colorado.edu/~palen/talks.html

[30] T.-Y. Liu, "Learning to Rank for information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[31] T. Joachims, "Training linear SVMs in linear time," in *Proc. ACM SIGKDD*, 2006, pp. 217–226.

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS*, 2013, pp. 3111–3119.