Noname manuscript No.
(will be inserted by the editor)

Ranking & Grouping Social Media Requests for Emergency Services Using Serviceability Model

Hemant Purohit · Carlos Castillo · Rahul Pandey

Received: date / Accepted: date

Abstract Social media has become an alternative communication mechanism for the public to reach out to emergency services during time-sensitive events. However, the information overload of social media experienced by these services, coupled with their limited human resources, challenges them to timely identify, prioritize, and organize critical requests for help. In this paper, we first present a formal model of serviceability called *Social-EOC*, which describes the elements of a *serviceable* message posted in social media expressing a request. Using the serviceability model, we then describe a system for the discovery and ranking of highly serviceable requests as well as for re-ranking requests by semantic grouping to reduce redundancy and facilitate the browsing of requests by responders. We validate the model for emergency services by experimenting with six crisis event datasets and ground truth provided by emergency professionals. Our experiments demonstrate that features based on both serviceability model as well as social connectedness improve the performance of discovering and ranking ($nDCG$ gain up to 25%) service requests over different baselines. We also empirically validate the existence of redundancy and semantic coherence among the serviceable requests using our semantic grouping approach, which shows the significance and need for grouping similar requests to save the time of emergency services. Thus, an application of serviceability model could reduce cognitive load on emergency servicers in filtering, ranking, and organizing public requests on social media at scale.

Purohit thanks US National Science Foundation grants IIS-1657379 & 1815459, and Castillo thanks La Caixa project LCF/PR/PR16/11110009 for partial support to this research.

H. Purohit *(corresponding author)*, R. Pandey
George Mason University, 4400 University Dr, MS 1G8, Fairfax, USA
Tel.: +1-703-993-5838
E-mail: {hpurohit,rpandey4}@gmu.edu

C. Castillo
Universitat Pompeu Fabra, Barcelona, Spain
E-mail: carlos.castillo@upf.edu

Keywords Information Overload · Serviceability · Social Media · Emergency Management · Semantic Grouping

1 Introduction

Recent years have shown the significant integration of social media into our daily life activities. This trend indicates the potential role of social media to timely connect the public to all kinds of organizations, including governments and for-profits. In for-profit companies, recent years have demonstrated the value of extending their customer relationship services to social media [Kietzmann et al. 2011], where they often provide timely answers to social media queries from existing and potential customers. Similarly, the research focused on public sector has shown that the public expects a timely response to queries on social media addressed to governments and nonprofits [American Red Cross 2012, Hughes et al. 2014, Reuter and Spielhofer 2017].

From the perspective of emergency services, however, there are substantial challenges for meeting these public expectations of timely response during emergencies. There are vast amounts of messages posted in social media by the public during emergencies [Yadav and Rahman 2016, Castillo 2016], leading to information overload in emergency services [Yin et al. 2012, Imran et al. 2015, Kibanov et al. 2017, Interdonato et al. 2019], 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 [Purohit et al. 2013, He et al. 2017, Ranganath et al. 2017] to damage reports [Madichetty and Sridevi 2019] and unsubstantiated rumors [Starbird et al. 2014]. Furthermore, messages often have syntactic and semantic-level redundancies in the content as well. Thus, quickly prioritizing and grouping redundant messages with help-seeking intent that require a timely response has become a critical need for agencies in the emergency operation centers (EOCs) [U.S. Homeland Security 2014].

Table 1 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

Table 1: 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*

(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. Therefore, we address the problem of filtering, prioritizing, and grouping serviceable social media requests for emergency services semantically and at scale.

Our contribution. To the best of our knowledge, this is the first comprehensive study to formally define and extensively analyze the application of a generic *serviceability* model (*Social-EOC*) for social media requests, in order to identify, prioritize, and group actionable requests to respond for emergency services. This paper builds upon our previous work [Purohit et al. 2018], where we proposed the *Social-EOC* model and presented experiments to rank social media requests using a Learning-to-Rank framework.

In contrast to our prior work, we investigate redundancy at the syntactic and semantic levels of requests that may potentially waste efforts of emergency responders and propose a semantic grouping approach to meaningfully exploit redundancy. We extend our earlier work by proposing a re-ranking method for the clusters of serviceable requests based on their semantic grouping using fine- and coarse-granularity clustering. This method provides more efficient interaction with public requests at scale and allows a timely response. Furthermore, we explore the role of a 'sociability' feature (c.f. details in Section 4.4) capturing social network characteristics of the requesting user in the ranking of serviceable requests. We investigate if the user's credibility manifested through social connectedness could contribute toward higher ranking of serviceable requests (*c.f.* analysis of sociability feature on ranking performance in Section 7.1).

We introduce the serviceability model in Section 3 together with a qualitative and quantitative description of serviceability characteristics of user requests. A learning-to-rank system based on the proposed model is described in Section 4, using inferences of the serviceability characteristics. We present the cluster-aware re-ranking methodology in Section 5. This system classifies and ranks serviceable requests while systematically exploiting content redundancy to organize requests by semantic grouping, for ultimately reducing the cognitive load on EOC personnel in processing public requests at large scale.

Finally, we demonstrate the validity of the ranking and grouping methodologies using the *Social-EOC* model in Sections 6 and 7, by experimenting with real-world datasets from six crisis events during the past six years. Our conclusions as well as lessons for future research appear in Section 8. The application of this research can ultimately help reduce the cognitive load on emergency service personnel in processing public requests on social media at large scale, and similarly, it can inform the research on improving social media services for customer care of various businesses as well.

2 Related Work

This section describes closely related work on the topic of filtering and ranking social media data for emergency management (for a survey, see [Imran et al. 2015, Castillo 2016]), primarily focusing on the research done to determine what is serviceable (or even relevant and actionable) for emergency services. Additionally, we summarize the related literature on methods to group ranked results of web-based retrieval systems.

2.1 Social media during emergencies

"Big Crisis Data" from social media has such a high volume, variety, and velocity, that it can overwhelm response services [Castillo 2016]. Crisis informatics [Palen and Anderson 2016] 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 [Imran et al. 2015]. Prior research has identified information overload as a key challenge and a barrier for the efficient use of social media communication by emergency services [Hiltz et al. 2014, Castillo 2016]. 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.

2.2 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 [Hughes and Palen 2012]. PIOs are provided guidelines for communication with the public [FEMA 2017], 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 [U.S. Homeland Security 2014, U.S. Homeland Security 2016, Reuter and Spielhofer 2017] 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.

2.3 Mining intent of requesting help

The literature provides some guidance on modeling requesting behavior or information seeking intent across different domains [Mai 2016], including Q&A

forum [Vasilescu et al. 2014], email communication [Yang et al. 2017], and social media platforms [Ferrario et al. 2012, Sachdeva and Kumaraguru 2017, Purohit et al. 2013]. 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 [Yang et al. 2017], 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 [Sachdeva and Kumaraguru 2017] to implicit requests for seeking donations and resources [Purohit et al. 2013, Varga et al. 2013] and other actions [Zade et al. 2018, Ranganath et al. 2017] during disasters. In particular for explicit requests to agencies, Sachdeva et al. [Sachdeva and Kumaraguru 2017] 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. Likewise, Ferrario et al. [Ferrario et al. 2012] 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 [Purohit et al. 2013, Nazer et al. 2016, He et al. 2017] proposed methods to identify implicit request messages for seeking or offering help during disaster relief, however, not specifically targeted to emergency services. Ranganath et al. [Ranganath et al. 2017] 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. Recently Zade et al. [Zade et al. 2018] presented a survey-based study with practitioners on defining actionability of social media messages during disaster events. Although relevant to our concept of serviceability, the actionability of content is considered generically and thus, it is not specific to requesting help and seeking response by the emergency services. 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 emergency services.

2.4 Grouping ranked results

Prior research in Information Retrieval has explored various ways to organize search ranking results and enable users to better interact with information and to reduce the cognitive effort of scanning through a long list [Wang and Zhai 2007, Osiński and Weiss 2005]. *Carrot*^{2} [1] is an example of a retrieval system for

[1] `https://project.carrot2.org/`

Table 2: 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 (10/28-11/06)	1,153	60	24 (40%)	36 (60%)
Oklahoma Tornado 2013 (05/20-06/10)	1,513	52	25 (48%)	27 (52%)
Louisiana Floods 2016 (08/14-09/29)	1,369	56	19 (34%)	37 (66%)
Alberta Floods 2013 (06/21-07/05)	2,727	814	229 (28%)	585 (72%)
Nepal Earthquake 2015 (04/25-05/11)	2,222	240	43 (18%)	197 (82%)
Hurricane Harvey 2017 (08/29-09/20)	12,742	1,534	306 (20%)	1,228 (80%)

web search that groups web results under different topical categories [Osiński and Weiss 2005]. Recent works [Wasilewski and Hurley 2016, Cobos et al. 2014] have continued exploring semantic grouping of ranked items for the different ways of incorporating diversity in ranking. However, to the best of our knowledge previous work has not investigated the grouping and ranking of short, informal language text of social media messages, especially in a time-sensitive scenario like disaster events.

3 *Social-EOC*: Serviceability Model

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

3.1 Qualitative serviceability model

We consider a general class of emergency service requests, following official guidelines from the US FEMA (Federal Emergency Management Agency) [FEMA 2017], 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, 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.

3.2 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 [Likert 1932]:

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 1.
- 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 1.

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 1.

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), subtype of resource (which), to make the request or question unambiguous, e.g., message *M1* in Table 1.

Our quantitative serviceability model defines a message m as a function of these characteristics $f(E(m), A(m), C(m), D(m))$. In this study, we automatically learn this scoring function instead of manually providing it, by estimating a relevance function learned via a Learning-to-Rank [Liu 2009] algorithm. We describe its implementation to rank messages next.

4 Ranking Requests using *Social-EOC* Serviceability Model

The proposed system summarized in Figure 1, 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, and
D) learning to classify and rank serviceable requests.

We present details of each of these steps in the following.

4.1 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 the CitizenHelper system [Karuna et al. 2017] and for prior events, we re-used datasets available from previous works [Sheth et al. 2014, Imran et al. 2016]. Following the recommendation of collecting "contextual streams" [Palen 2014], 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 2 shows a summary of the characteristics of our dataset.

4.2 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

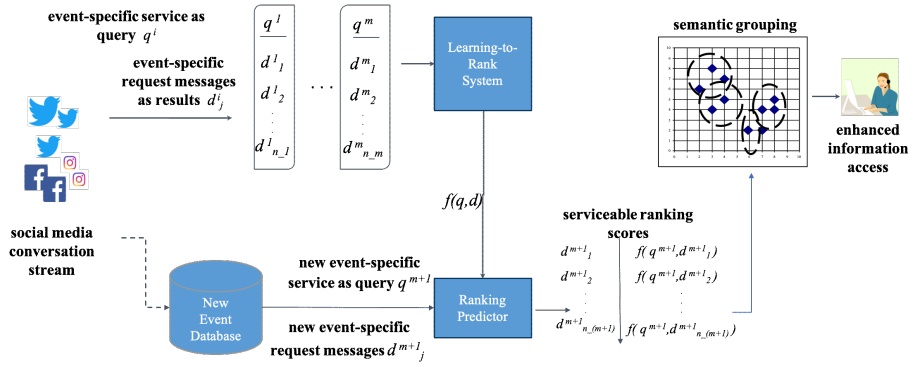

Fig. 1: System design based on *Social-EOC* model and using the Learning-to-Rank approach for prioritization and fine/coarse granularity clustering for semantic grouping.

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 3. 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, 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*

Table 3: 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

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 1 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 3 shows examples of messages with average ratings.

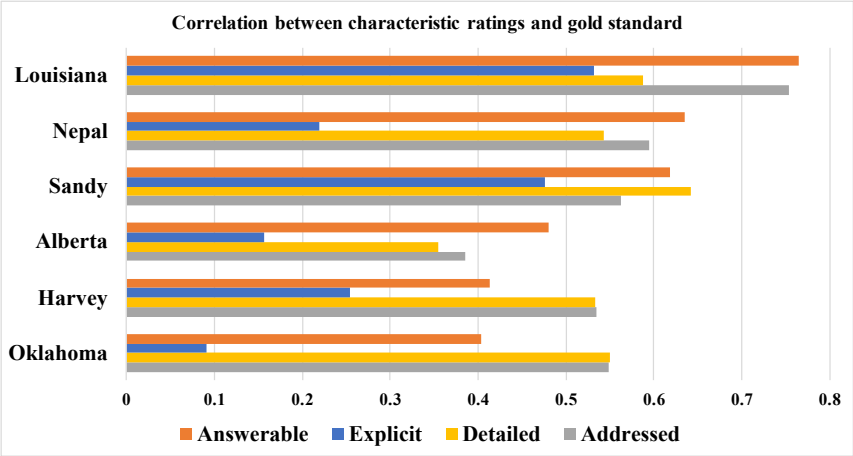

Fig. 2: We observe a positive correlation between the average crowdsourced (non-expert) ratings of the proposed serviceability characteristics and the gold standard annotations of serviceability given by domain experts, according to Pearson Correlation. All correlations are statistically significant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

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

Table 2 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 3.

Second, a disagreement between the experts was observed in relation to messages that only express gratitude (such as M3 in Table 1). One of the three experts considered such messages as serviceable, with the rationale that replying to gratitude messages could help strengthen trust within a commu-

nity. 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 to counter 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.

4.4 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 [Liu 2009] 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 $D^i = \{d_1^i, d_2^i, \ldots, d_{n\_i}^i\}$ 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_j^i.

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:

$$f_i(u, v) = \left\{ \begin{array}{l} -1, \ (Y^i)_u > (Y^i)_v \\ 1, \quad (Y^i)_v > (Y^i)_u \end{array} \right\} \tag{1}$$

To solve this problem, we use the SVM-Rank algorithm [Joachims 2006].

Feature Extraction. Query-document pairs (q^i, d_j^i) 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_`.
- *Social features*: A measure to capture the degree of sociability or connectedness of a user in the social network who posts a request message. We hypothesize that the users with higher social connectedness are less likely to tweet non-serviceable request messages to emergency services, given the potential impact on their credibility. Social connectedness has been shown to associate with the credible perception of users, as studied in the prior research [Westerman et al. 2012, Riedl et al. 2013]. Thus, we consider its exploration in our study, whether it could contribute toward higher ranking of serviceable requests sourced from users with potentially high credibility. In this study, we define a very simple measure of the connectedness of a user in the network based on both in-degree (number of friends) and out-degree (number of followers) metadata available with the Twitter user profile. It is calculated as follows:

$$Sociability = \log\left(1 + \frac{1 + friends\_count}{1 + followers\_count}\right) \tag{2}$$

- *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[Mikolov et al. 2013]. 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.

5 Re-ranking Requests by Semantic Grouping

The existence of redundancy at both syntactic (near duplicity) and semantic (near topicality) levels among the ranked requests, translate into inefficiencies for emergency servicers, who have to browse through multiple messages that communicate either the same or similar requests. Therefore, we propose to cluster the results of serviceability ranking of individual requests, creating semantic groupings and re-ranking individual requests by clusters as described next.

5.1 Grouping types.

We create two types of clusters of the individual ranked requests:

- ***Fine-Granularity Clusters***. This type of clustering aims to group together near-duplicates of request messages for a sub-event, so that an emergency servicer can choose to respond to the desired set of similar event-related requests simultaneously. Such grouping organically clusters syntactically similar messages in a bottom-up approach. For example, social media requests related to a specific charity that is accepting donations during a hurricane would have similar messaging content and can be grouped together.
- ***Coarse-Granularity Clusters***. This type of clustering aims to group request messages by their topical functionality as described by the topical categories in a structured domain ontology. Such domain ontology provides a top-down approach to organize and structure information in the existing information management systems in a domain. Thus, it provides a seamless, efficient way to cluster and browse the topically-similar ranked requests for an emergency servicer, for example, the social media requests related to transportation topic category. Multiple ontologies (e.g., HXL [Keßler and Hendrix 2015], MOAC [Limbu et al. 2012]) have been proposed for the emergency management domain to support disaster information management systems such as SoKNOS [Babitski et al. 2011]. However, they tend to be too fine-grained for our purpose of grouping requests. Instead, we demonstrate the proof-of-concept for the content browsing based on semantic grouping of ranked requests using the DBpedia ontology[Auer et al. 2007], which can be mapped to a well-known taxonomy used by practitioners, as explained in the next section.

5.2 Feature representation and clustering.

The fine-granularity clustering method represents each request message using the bag-of-words model after pre-processing text of requests. We use *TF-IDF* weighing scheme for a word feature and compute cosine distance between two word vectors of request messages for the clustering process [Baeza-Yates et al. 2011]. Our pre-processing replaces URLs, numbers, and user mentions by generic tokens, and removes stopwords using a dictionary plus the top 3% frequent words in the vocabulary. We used hierarchical clustering with no predefined criterion on the desired number of clusters.

We perform the coarse-granularity clustering using a distributed semantic representation of words as vectors (a word embedding), specifically, the pre-trained embeddings of Wikipedia concepts and entities (e.g., *ConVec* [Sherkat and Milios 2017].) We compute the cosine distance-based similarity to match the embedding representations of a request message (averaging over message tokens' embeddings) and the concept of the relevant top-level category from

Table 4: Illustration of ontological categories from DBpedia, used for the semantic grouping under the coarse-granularity clustering type.

DBpedia Class	Wikipedia Page Title	Description
Medicine	Medicine	Healthcare related requests, with subcategories
Food	Food	Basic logistics related queries
MeanOfTransportation	Mode of transport	Requests on road, rail or air infrastructure
Population	Population	Affected people related queries
PublicService	Public service	Requests related to functions of services

the DBpedia ontology as shown in Table 4. Thus, this approach can facilitate a top-down faceted browsing of requests under various ontological subcategories. We determine a request item's membership to a concept-category based cluster by the highest matched concept category. Relevant concept categories such as *Medicine* or *MeanOfTransportation* are determined based on the framework of well-known Incident Command System (ICS) [FEMA 2017] used by emergency management practitioners. We consider the set of requests under the same concept category as a cluster, which can have sub-clusters.

The advantage of the resulting clustered messages using such hierarchical approach is that we can further navigate the top-level concept cluster for sub-categories in the hierarchy of the predefined domain ontology, or the dendogram of the fine-granularity clusters. The flexibility for faceted navigation is useful for emergency responders with different roles for response (e.g., officers responding to transportation queries and within that, road versus rail transport).

5.3 Re-ranking.

For each of the clustering approaches, we retrieve clusters that have different sizes and that involve messages with different relevance scores from the individual ranking method described in the previous section. Our aim is to rank the clusters, where the cluster importance is measured using the relevance rank of individual request messages contained in the cluster.

The importance function denoted by $score\_rank(c_k)$ for a cluster c_k is computed by following *Borda Count* [Yang and Guo 2016], which is a preference aggregation method where scores are given to each candidate cluster in reverse proportion to their ranking. Thus, the higher-ranked candidates receive more points and we sort candidate clusters by $score\_rank(c_k)$ values, defined as:

$$\text{score\_rank}(c_k) = \text{MAX\_RANK} - \text{rank}(c_k) \tag{3}$$

where c_k refers to k^{th} cluster with individual request members d^i_j, MAX_RANK denotes the maximum rank a cluster could achieve (i.e. the total number of clusters), and $rank(c_k)$ function provides the representative position of the cluster c_k among the candidate clusters. This is estimated

by ordering the best relevance scored member (representative) of each of the candidate clusters (i.e. the member d_j^i with the highest relevance score in the individual serviceability ranking among the request members of c_k).

6 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 individual requests. In all cases, generic features are computed and text is preprocessed as described in Section 4.4.

- [***T***]: **Text + generic features** (baseline). This method uses a standard bag-of-words (BoW) representation of the text features, along with the generic features.
- [***T +S***]: **Text + generic features + social**. This method uses features from baseline scheme T plus social 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 +S***]: **T + inferred labels + social**. This method uses features from T+I scheme plus social features.
- [***T + I_all***]: **T + inferred from all-events model**. This method uses features from T plus serviceability characteristic labels generated by a classifier trained on a held-out portion of messages from all 6 events.
- [***T + I_all +S***]: **T + inferred from all-events model + social**. This method is similar to the preceding scheme, plus social features.
- [***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 + I_cross +S***]: **T + inferred from cross-event model + social**. This method is similar to the preceding scheme plus social features.
- [***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_cross + I_cross +S***]: **T cross-events + inferred from cross-event model + social**. This method uses social features in addition the preceding scheme.
- [***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 would 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 ($nDCG$) [Liu 2009], which effectively compares two rankings by weighing more differences in the top positions than differences further down. Specifically, for each event/query:

$$nDCG(k) = G^{-1}_{max,i}(k) \sum_{j:\pi_i(j)\leq k} \frac{2^{y_{i,j}} - 1}{\log_2(1 + \pi_i(j))} \tag{4}$$

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

We analyzed $nDCG$ for the top-5 and top-10 items, for the rankings obtained across each fold of the 5-fold cross validation setting, for each event.

Re-ranking implementation. For the re-ranking approach based on semantic grouping of requests, we use the $nDCG$ metric for assessing the quality of the ranking of clusters. Naturally, in this case the relevance is computed at the cluster level as describe earlier. We consider the cluster representative member's relevance score as the ground truth for evaluation. Given the better performance across majority of cases, we chose the output of T + I_cross + S classification and ranking model scheme for the input to the re-ranking method. We have taken the top-200 ranked messages by relevance for our two clustering approaches. For implementing fine-granularity clustering, we used Scipy's Hierarchical Clustering algorithm with no predefined cluster size and taking the similarity threshold of 0.7 for all events except the smaller dataset of Louisiana (chose 0.55, in order to aggregate in greater than one cluster, for any meaningful analysis). We added a constraint to the cluster size for containing minimum 1 item, in case of highly unrelated request messages. Furthermore, for coarse-granularity clustering, we used the publicly available vector embeddings of Wikipedia Concepts and entities, called $ConVec$ [Sherkat and Milios 2017] (`https://github.com/ehsansherkat/ConVec`).

7 Results and Discussion

In this section we first present the results from the individual request ranking schemes, comparing their performance with respect to the diverse features of the serviceability model. We then present the results of the re-ranking for the semantically grouped requests.

7.1 Result observations for individual request ranking

Table 5 compares the performance of different ranking model schemes described in section 6, in terms of $nDCG$ values of the first 5 positions ($nDCG$@5)

Table 5: Comparison of *nDCG*@5 and *nDCG*@10 (expressed as percentages) using 5-fold cross validation for each event. (**Small dataset with no. of labeled instances < 100*)

Event	**Classification Schemes**	*nDCG***@5**	*nDCG***@10**
Oklahoma*	T *(baseline)*	49%	74%
	T+S	50%	75%
	T+I	53%	77%
	T+I+S	**59**%	**83**%
	T+I_all	46%	72%
	T+I_all+S	48%	72%
	T+I_cross	42%	71%
	T+I_cross+S	58%	82%
	T_cross+I_cross	46%	72%
	T_cross+I_cross+S	40%	70%
	T+M *(hand-labeled)*	*61%*	*85%*
Louisiana*	T (baseline)	94%	96%
	T+S	87%	95%
	T+I	89%	96%
	T+I+S	89%	95%
	T+I_all	89%	96%
	T+I_all+S	89%	95%
	T+I_cross	**96**%	**99**%
	T+I_cross+S	92%	98%
	T_cross+I_cross	77%	90%
	T_cross+I_cross+S	83%	94%
	T+M *(hand-labeled)*	*93%*	*98%*
Sandy*	T *(baseline)*	50%	67%
	T+S	50%	68%
	T+I	57%	75%
	T+I+S	57%	75%
	T+I_all	57%	75%
	T+I_all+S	57%	75%
	T+I_cross	**71**%	**87**%
	T+I_cross+S	67%	85%
	T_cross+I_cross	56%	79%
	T_cross+I_cross+S	51%	75%
	T+M *(hand-labeled)*	*72%*	*90%*
Nepal	T (baseline)	46%	44%
	T+S	46%	43%
	T+I	52%	50%
	T+I+S	50%	48%
	T+I_all	55%	50%
	T+I_all+S	55%	50%
	T+I_cross	52%	50%
	T+I_cross+S	50%	48%
	T_cross+I_cross	58%	63%
	T_cross+I_cross+S	**61**%	**64**%
	T+M *(hand-labeled)*	*74%*	*66%*
Alberta	T (baseline)	57%	47%
	T+S	55%	54%
	T+I	56%	52%
	T+I+S	54%	56%
	T+I_all	49%	53%
	T+I_all+S	55%	**65**%
	T+I_cross	56%	52%
	T+I_cross+S	54%	56%
	T_cross+I_cross	65%	59%
	T_cross+I_cross+S	**66**%	61%
	T+M *(hand-labeled)*	*91%*	*84%*
Harvey	T (baseline)	62%	60%
	T+S	59%	56%
	T+I	**64**%	62%
	T+I+S	64%	59%
	T+I_all	62%	**69**%
	T+I_all+S	64%	64%
	T+I_cross	64%	62%
	T+I_cross+S	64%	59%
	T_cross+I_cross	54%	52%
	T_cross+I_cross+S	54%	54%
	T+M *(hand-labeled)*	*74%*	*78%*

and 10 positions ($nDCG$@10) of ranking. For a qualitative analysis of the resulting ranking systems, we also present top-2 and bottom-2 ranked items from ***T+I+S*** method in table 6. We observe the following from these tables:

1**.) Social features are useful for ranking serviceable requests.** Features based on social network characteristics help in finding and ranking serviceable request messages (see ***T+I+S*** results and other model variants with ***S*** in table 5). Although, our experimental results showed some dependence on the dataset size for the effectiveness of this feature type; for the majority of the cases, the results show a favorable pattern over the baseline. Thus, we recommend to use social features but also suggest testing them with respect to specific deployment of serviceability ranking systems.

2**.) The serviceability characteristics of our model capture the notion of serviceability desired by domain practitioners.** The performance of the ranking model schemes based on inferring serviceability characteristics is above the performance of the baseline in all cases, and if serviceability characteristics are given as manual inputs (i.e. the method ***T+M***, *hand-labeled*), we obtain the best performance (except in the case of Louisiana due to very small labeled dataset for training, we explain it further in the observation 4.) 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 and our proposed ranking model schemes are advantageous.

3**.) Inferring serviceability characteristics is better than the baseline.** Overall, there is a consistent pattern of good performance across the proposed ranking model schemes (i.e. ***T+I*** and variants). The improvement in $nDCG$@5 and $nDCG$@10 is obtained by adding the proposed serviceable characteristics features (automatically inferred) as well as the social features to the baseline text (bag-of-words) features (***T***), thus, demonstrating the significance of these features for ranking highly serviceable requests quickly and efficiently in general.

4**.) Ranking performance varies in the cases of small datasets.** Among the small datasets, we observe that in the Louisiana event, for top-5 positions (***T+I***) results are not better than ***T***. Although ***T+I_cross*** result is better, which has an advantage to leverage other large event datasets for training an efficient model. 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. There are less training examples for the small datasets (i.e. Oklahoma, Louisiana, Sandy) that limit the performance of a learning-to-rank model when using only the event-specific data, which is in line with our next observations.

5**.) Cross-event models perform well.** We found promising results for majority of the ranking model schemes 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 (see ***T_cross + I_cross***) for the Harvey event with large dataset, given that in

the Harvey dataset we have more labeled data than all other events combined. Thus, the performance of this cross-event modeling scheme with less training data to leverage deteriorates for Harvey event.

6.) **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 discriminatory 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 5.

7.) **The ranked items from the resulting ranking system match the expectations on the qualitative serviceability characteristics.** Examples in table 6 for the top ("best") and bottom ("worst") ranked messages shows good performance in prioritization, using the ***T+I+S*** method based on text (bag-of-word) as well as inferred serviceability characteristics and social features. The examples match the expectations on the qualitative serviceability characteristics for all of the cases, by prioritizing important messages among the top items. For instance, in the first case of Hurricane Sandy event, the top-ranked messages clearly have specific queries that could be answered by emergency services, while the bottom ones are insignificant for disaster response and are likely to waste the time of emergency services.

7.2 Result observations for grouped requests ranking

We first study the distribution of resulting clusters from our two methods of fine- and coarse-granularity in figures 3 and 4, followed by the analysis of cluster-wise group ranking performance in figure 5 and table 7. We note the following observations:

1.) **Fine-grained grouping method primarily shows the presence of incoherent clusters.** The histograms of cluster-size in figure 3 show equal-sized clusters across the events (except Louisiana and Nepal with some dense clusters as well) that indicate the presence of syntactic redundancy in social media requests. These varying cluster sizes validate our initial motivation to systematically exploit the behavior of redundancy of requests for the ranking.

2.) **Coarse-grained grouping method results in highly imbalanced clusters.** Figure 4 shows the consistency in the imbalanced cluster distribution of social media requests across all the events. Note that the small datasets have requests of only few types, yet with the majority of the instances belonging to the specific common category of DBpedia class 'PublicService' that is present across all the events. Although both sparse and dense clusters of serviceable requests can result from this method, there is higher coherence among the requests. This indicates the possibility to design a top-down faceted browsing system for semantically grouped social media requests across the subcategories of a top category using DBpedia ontology.

Table 6: Top-2 (most serviceable) and bottom-2 (least serviceable) tweets for each event across the last decade, obtained automatically using the ***T+I+S*** ranking model scheme.

Event	**Ranked Requests**
Hurricane Sandy 2012	
[TOP]	* Queens trains aren't being addressed at all. When can we expect any service updates for the NQR trains? Please advise! * please, governor, post a website or phone # where we can get specific info for our local areas
[BOTTOM]	- Romney's not going 2 like that Gov Christie is being nice about Obama's leadership during Sandy. His pittbull is playing nice. - HILARIOUS! That is much needed laughter, I am sure.
Oklahoma Tornado 2013	
[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]	- I agree with *ANONYMOUS*. Any answers? - help us too!!! #oklahoma
Alberta Floods 2013	
[TOP]	* can you tell me if sanitary pumps are running yet in elbow park? #yycflood * Can you point us to where we can get a prompt tetanus booster shot? Some of us had submerged cuts and nicks. #yycflood
[BOTTOM]	- thank u calgary police - Thank for your time.
Nepal Earthquake 2015	
[TOP]	* cc @account0 have a collection point here in Hyd. Let me see if I can tag u to another tweet. * @account0 able to organise a collection of goods you mention but can Goonj guarantee they have capacity on ground to deliver?
[BOTTOM]	- send 100 bits to @account1 for #NepalEarthquake disaster recovery. - I'm trying to send a fiver, FFS...
Louisiana Floods 2016	
[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 ?
Hurricane Harvey 2017	
[TOP]	* This list of neighborhoods has caused confusion, we need clarity on exactly what areas are impacted by this. * Can you please help? #HarveySOS #Harvey #AnimalRescue #HumanRescue #DisabledRescue URL
[BOTTOM]	- I did thank you! - Yes thank you #harvey

3.) **Cluster rank is negatively correlated with the cluster size.** Figure 5 shows the Pearson correlation coefficient between the cluster size versus the cluster rank in case of both types of clustering. This indicates the higher ranked (more relevant) requests tend to be topically unique, while the lower-ranked (less relevant) requests have greater levels of content redundancy. This also suggests there are some messages that have relatively low relevance and are repeated much more in social media than messages that are relatively more important. In other words, redundancy alone is a weak signal of relevance.

4.) **Ranking of the clusters of requests using both fine and coarse approaches shows good performance, based on $nDCG@5$ metric.** As shown in table 7, re-ranking of grouped/clustered requests shows good ranking performance with average $nDCG@5$ greater than 98% for the ranked clusters from both types of clustering across all events. This suggests that the

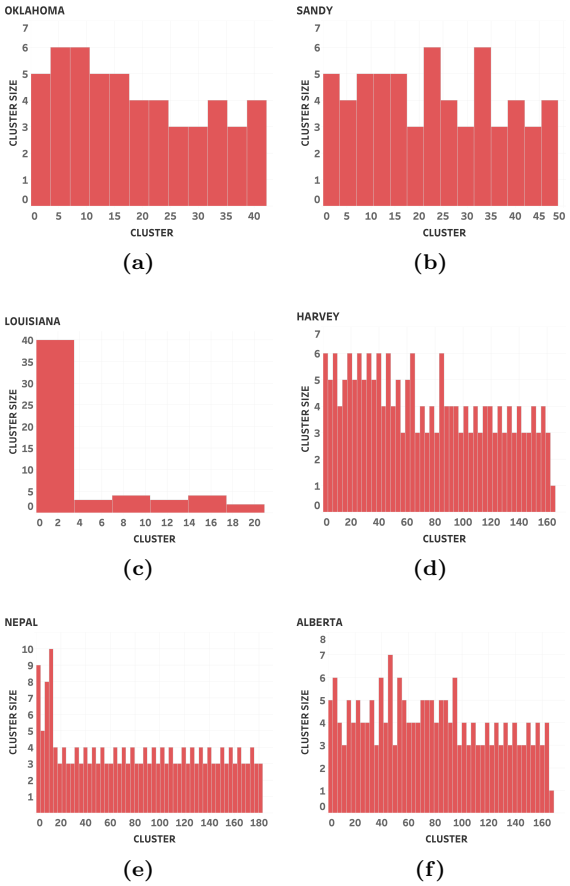

Fig. 3: Histogram of cluster size distribution across events for the results from fine-granularity clustering method.

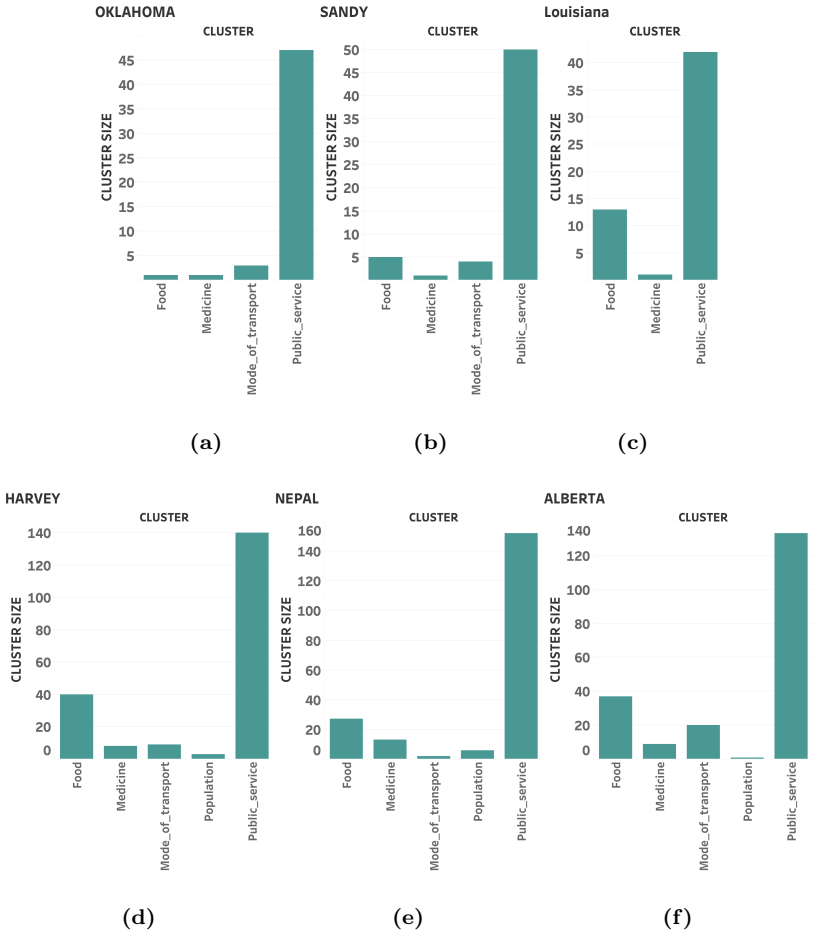

Fig. 4: Histogram of cluster size distribution across events for the results from coarse-granularity clustering method that shows the cluster label as the DBpedia category.

Table 7: Comparison of $nDCG$@5 (expressed as percentages) for the cluster ranking performance for each event using both fine- and coarse-granularity. (*chose k=5 for consistency across both clustering types*)

Event	$nDCG$**@5** Fine-granularity	$nDCG$**@5** Coarse-granularity
Oklahoma	100%	100%
Sandy	100%	100%
Louisiana	100%	100%
Nepal	100%	100%
Harvey	100%	96.75%
Alberta	86.88%	100%

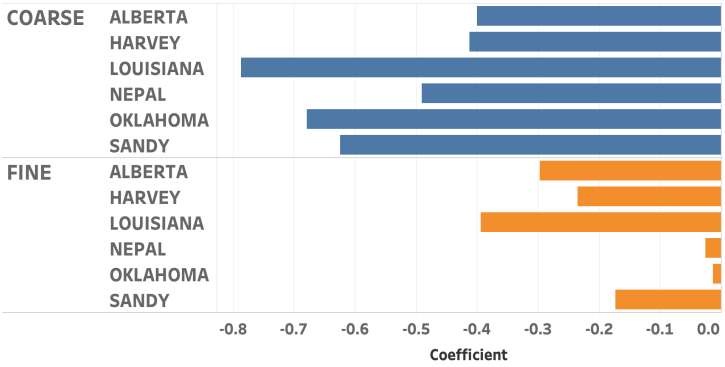

Fig. 5: Pearson correlation coefficients between the cluster rank and cluster size for the resulting clusters of semantic grouping across all events, indicating uniqueness of the higher-ranked requests.

re-ranking by semantic grouping may not impact the prior ranking system's performance metrics and the proposed grouping approach rather acts as a filter for efficient interaction with semantically similar information.

5.) **Grouping social media requests using semantic representation of coarse-granularity method is useful for long-term disaster response.** Given the coarse-granularity clustering results into more coherent clusters, it is a better approach for the longer periodic grouping of requests, where there is greater likelihood of multiple coherent topics of requests instead of only exact or near duplicates. In contrast, fine-granularity clustering method is useful for grouping and ranking in the short time intervals, as it groups and reduces near-duplicates of same content with higher syntactic overlaps.

8 Conclusions and Future Work

This paper presents a novel study of serviceable request characteristics of social media messages, which can improve prioritization, filtering, and organization of request messages from social media streams for emergency services. This is done through a novel model for serviceability of social media requests sent to an organization or agency by a user, called the *Social-EOC* serviceability model. We should note that in some cases serviceable requests are posted by users directly affected, who request help for themselves. In other cases, those affected may not be able to post a message and the serviceable message may be posted on their behalf by somebody else.

We demonstrated the applicability of this serviceability model for emergency services by creating different types of classification and ranking systems using the proposed serviceability characteristics of a request message, social network characteristics of connectedness of the requesting user, as well as a novel semantic grouping method for addressing redundancy of the ranked requests. Specifically, we proposed several systems for classifying and ranking

requests for serviceability, with a baseline text-based method using bag-of-words features, and a series of variants of our method while leveraging inferred features for the serviceability characteristics and social features. We further presented a methods for re-ranking and semantic grouping of requests by two clustering types of fine- and coarse-granularity.

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%). We also observed the effectiveness of the proposed semantic grouping method in reducing the redundancy of similar requests yet preserving the high relevance requests in the clusters at the top of cluster ranking, as observed by the average $nDCG@5$ above 98% for the cluster rankings across events. 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 often get overwhelmed during mass emergencies.

Limitations and future work. While our presented method can prioritize and group serviceable messages, it has some limitations.

First, the analyzed data contains only English language messages, which is the dominant language in the data 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 dataset come from the same platform, i.e. Twitter, and we would need to evaluate messages in other platforms where populations and norms may be different and hence, serviceability might differ.

Second, 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 $nDCG$ measure can be used with non-binary relevance assessments.

Third, we focused on the semantic grouping of requests based on textual content, while there can be also grouping by spatial and temporal semantics. We plan to explore the ensemble of diverse semantic grouping approaches in a future study.

Fourth, there is a possibility of temporal significance of relevancy for the posting time when a request was posted, such as when the hurricane landfall occurs. However, in this research, the data collection was focused on the period of active response to the emergency. This translates to the observation from our temporal analysis that the percentage of relevant messages per day never fell below a certain threshold in an event. Thus, a system based on this method would require some pre-filtering of relevant messages, especially if applied outside the active response period. A future work could explore this direction further for temporal significance of requests to develop a generic social media filtering and ranking system for all phases of the disaster management.

Fifth, there is no availability of realtime data for the actual social network structure, which could be leveraged for generating more features based on social network characteristics to benefit the real-time serviceability ranking applications.

Sixth, there is a likelihood of a user sending emotional messages for requesting help, or using certain patterns of vocabulary that is used by people under extreme stress and can be a part of conversational context. Therefore, features capturing such subjectivity could be explored for ranking serviceable requests in the future work.

Reproducibility. Anonymized messages, a list of official accounts, the crowdsourced and expert labels as well as codes will be available with the accepted version of this manuscript.

9 Acknowledgement

We thank *(to be added after acceptance)*

References

American Red Cross 2012. American Red Cross (2012) More americans using mobile apps in emergencies. `http://www.redcross.org/news/press-release/More-Americans-Using-Mobile-Apps-in-Emergencies`, online and phone survey

Auer et al. 2007. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) Dbpedia: A nucleus for a web of open data. The semantic web, Springer, pp. 722–735

Babitski et al. 2011. Babitski G, Bergweiler S, Grebner O, Oberle D, Paulheim H, Probst F (2011) Soknos–using semantic technologies in disaster management software. In: Extended Semantic Web Conference, Springer, pp 183–197

Baeza-Yates et al. 2011. Baeza-Yates R, Ribeiro BdAN, et al. (2011) Modern information retrieval (Second Edition). New York: ACM Press; Harlow, England: Addison-Wesley

Castillo 2016. Castillo C (2016) Big Crisis Data: Social Media in Disasters and Time-Critical Situations. Cambridge University Press

Cobos et al. 2014. Cobos C, Muñoz-Collazos H, Urbano-Muñoz R, Mendoza M, León E, Herrera-Viedma E (2014) Clustering of web search results based on the cuckoo search algorithm and balanced bayesian information criterion. Information Sciences, 281:248–264

FEMA 2017. FEMA (2017) Public information officer (PIO). `https://training.fema.gov/programs/pio/`

Ferrario et al. 2012. Ferrario MA, Simm W, Whittle J, Rayson P, Terzi M, Binner J (2012) Understanding actionable knowledge in social media: BBC question time and Twitter, a case study. In: Proc. ICWSM, pp 455–4458

He et al. 2017. He X, Lu D, Margolin D, Wang M, Idrissi SE, Lin YR (2017) The signals and noise: Actionable information in improvised social media channels during a disaster. In: Proc. ACM WebSci, pp 33–42

Hiltz et al. 2014. Hiltz SR, Kushma JA, Plotnick L (2014) Use of social media by us public sector emergency managers: Barriers and wish lists. In: Proc. ISCRAM, pp 602–611

Hughes and Palen 2012. Hughes AL, Palen L (2012) The evolving role of the public information officer: An examination of social media in emergency management. Journal of Homeland Security and Emergency Management 9(1)

Hughes et al. 2014. Hughes AL, St Denis LA, Palen L, Anderson KM (2014) Online public communications by police & fire services during the 2012 hurricane Sandy. In: Proc. ACM SIGCHI, pp 1505–1514

Imran et al. 2015. Imran M, Castillo C, Diaz F, Vieweg S (2015) Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) 47(4):67
Imran et al. 2016. Imran M, Mitra P, Castillo C (2016) Twitter as a lifeline: human-annotated Twitter corpora for NLP of crisis-related messages. In: Proc. LREC
Interdonato et al. 2019. Interdonato R, Guillaume J, Doucet A (2019) A lightweight and multilingual framework for crisis information extraction from Twitter data. Soc. Netw. Anal. Min. 9(1):65
Joachims 2006. Joachims T (2006) Training linear SVMs in linear time. In: Proc. ACM SIGKDD, pp 217–226
Karuna et al. 2017. Karuna P, Rana M, Purohit H (2017) Citizenhelper: A streaming analytics system to mine citizen and web data for humanitarian organizations. In: Proc. ICWSM, pp 729–730
Keßler and Hendrix 2015. Keßler C, Hendrix C (2015) The humanitarian exchange language: coordinating disaster response with semantic web technologies. Semantic Web 6(1):5–21
Kibanov et al. 2017. Kibanov M, Stumme G, Amin I, Lee JG (2017) Mining social media to inform peatland fire and haze disaster management. Soc. Netw. Anal. Min. 7(1):30
Kietzmann et al. 2011. Kietzmann JH, Hermkens K, McCarthy IP, Silvestre BS (2011) Social media? get serious! understanding the functional building blocks of social media. Business horizons 54(3):241–251
Likert 1932. Likert R (1932) A technique for the measurement of attitudes. Archives of psychology
Limbu et al. 2012. Limbu M, Wang D, Kauppinen T, Ortmann J (2012) Management of a crisis (moac) vocabulary specification. Online at http://observedchange com/moac/ns
Liu 2009. Liu TY (2009) Learning to Rank for information retrieval. Foundations and Trends® in Information Retrieval 3(3):225–331
Mai 2016. Mai JE (2016) Looking for information: A survey of research on information seeking, needs, and behavior. Emerald Group Publishing
Madichetty and Sridevi 2019. Madichetty S, Sridevi M (2019) Disaster damage assessment from the tweets using the combination of statistical features and informative words. Soc. Netw. Anal. Min. 9(1):42
Mikolov et al. 2013. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proc. NIPS, pp 3111–3119
Nazer et al. 2016. Nazer TH, Morstatter F, Dani H, Liu H (2016) Finding requests in social media for disaster relief. In: Proc. IEEE/ACM ASONAM, pp 1410–1413
Osiński and Weiss 2005. Osiński S, Weiss D (2005) Carrot 2: Design of a flexible and efficient web information retrieval framework. In: International atlantic web intelligence conference, Springer, pp 439–444
Palen 2014. Palen L (2014) Frontiers of crisis informatics. Computer Science Colloquia, University of Colorado, Boulder, `https://www.cs.colorado.edu/~palen/talks.html`
Palen and Anderson 2016. Palen L, Anderson KM (2016) Crisis informatics – new data for extraordinary times. Science 353(6296):224–225
Purohit et al. 2013. Purohit H, Castillo C, Diaz F, Sheth A, Meier P (2013) Emergency-relief coordination on social media: Automatically matching resource requests and offers. First Monday 19(1)
Purohit et al. 2018. Purohit H, Castillo C, Imran M, 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), IEEE, pp 119–126
Ranganath et al. 2017. Ranganath S, Wang S, Hu X, Tang J, Liu H (2017) Facilitating time critical information seeking in social media. IEEE Transactions on Knowledge and Data Engineering 29(10):2197–2209
Riedl et al. 2013. Riedl C, Kbler F, Goswami S, Krcmar H (2013) Tweeting to feel connected: A model for social connectedness in online social networks. International Journal of Human-Computer Interaction, 29(10):670-687
Reuter and Spielhofer 2017. Reuter C, Spielhofer T (2017) Towards social resilience: A quantitative and qualitative survey on citizens' perception of social media in emergencies in Europe. Technological Forecasting and Social Change 121:168–180

Sachdeva and Kumaraguru 2017. Sachdeva N, Kumaraguru P (2017) Call for service: Characterizing and modeling police response to serviceable requests on Facebook. In: Proc. ACM CSCW, pp 336–352

Sherkat and Milios 2017. Sherkat E, Milios E (2017) Vector embedding of wikipedia concepts and entities. In: Proc. NLDB, pp 418–428. DOI 10.1007/978-3-319-59569-6_50, https://doi.org/10.1007/978-3-319-59569-6_50

Sheth et al. 2014. Sheth A, Jadhav A, Kapanipathi P, Lu C, Purohit H, Smith GA, Wang W (2014) Twitris: A system for collective social intelligence. In: Encyclopedia of social network analysis and mining, Springer, pp 2240–2253

Starbird et al. 2014. Starbird K, Maddock J, Orand M, Achterman P, Mason RM (2014) Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. Proc iConference

U.S. Homeland Security 2014. US Homeland Security (2014) Using social media for enhanced situational awareness and decision support. https://www.dhs.gov/publication/using-social-media-enhanced-situational-awareness-decision-support

U.S. Homeland Security 2016. US Homeland Security (2016) From concept to reality: Operationalizing social media for preparedness, response and recovery. https://www.dhs.gov/publication/vsmwg-concept-reality

Varga et al. 2013. Varga I, Sano M, Torisawa K, Hashimoto C, Ohtake K, Kawai T, Oh JH, De Saeger S (2013) Aid is out there: Looking for help from tweets during a large scale disaster. In: Proc. ACL, vol 1, pp 1619–1629

Vasilescu et al. 2014. Vasilescu B, Serebrenik A, Devanbu P, Filkov V (2014) How social Q&A sites are changing knowledge sharing in open source software communities. In: Proc. ACM CSCW, pp 342–354

Wang and Zhai 2007. Wang X, Zhai C (2007) Learn from web search logs to organize search results. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 87–94

Wasilewski and Hurley 2016. Wasilewski J, Hurley N (2016) Incorporating diversity in a learning to rank recommender system. In: The Twenty-Ninth International Flairs Conference.

Westerman et al. 2012. Westerman D, Spence PR, Van Der Heide B (2012) A social network as information: The effect of system generated reports of connectedness on credibility on Twitter. Computers in Human Behavior 28(1):199-206

Yadav and Rahman 2016. Yadav M, Rahman Z (2016) The social role of social media: the case of Chennai rains-2015. Soc. Netw. Anal. Min. 6(1):101

Yang and Guo 2016. Yang Y, Guo J (2016) Exact algorithms for weighted and unweighted borda manipulation problems. Theoretical Computer Science 622:79–89

Yang et al. 2017. Yang L, Dumais ST, Benne PN, Awadallah AH (2017) Characterizing and predicting enterprise email reply behavior. In: Proc. ACM SIGIR, pp 235–244

Yin et al. 2012. Yin J, Lampert A, Cameron M, Robinson B, Power R (2012) Using social media to enhance emergency situation awareness. IEEE Intelligent Systems 27(6):52–59

Zade et al. 2018. Zade H, Shah K, Rangarajan V, Kshirsagar P, Imran M, Starbird K (2018) From situational awareness to actionability: Towards improving the utility of social media data for crisis response. Proc ACM Hum-Comput Interact 2(CSCW):195:1–195:18, DOI 10.1145/3274464, http://doi.acm.org/10.1145/3274464