Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

# Multilingual Serviceability Model for Detecting and Ranking Help Requests on Social Media during Disasters

**Fedor Vitiugin**[*1], **Hemant Purohit** [2]

[1] Web Science and Social Computing Research Group, Universitat Pompeu Fabra, Spain
[2] Humanitarian Informatics Lab, George Mason University, USA
fedor.vitiugin@upf.edu, hpurohit@gmu.edu

## Abstract

Social media users expect quick and high-quality responses from emergency services when seeking help. However, these organizations face difficulties in detecting and prioritizing critical requests due to the overwhelming amount of information on social media and their limited human resources to tackle it during mass emergencies or disaster events. The situation is exacerbated when users communicate in different or native languages, which can be expected during disasters. While recent studies have focused on characterizing and automatically detecting help requests on social media, they focused on non-behavioral features and monolingual data, primarily in English. Thus, a key gap exists in analyzing multilingual requests on social media for public services.

In this paper, we introduce a knowledge distillation framework called MulTMR (Multiple Teachers Model for detecting and Ranking), which combines the power of both task-related and behavior-guided models as diverse teachers for training a student model to efficiently detect serviceable request messages across languages and regions on social media during natural disaster events. We demonstrate that the presented framework can enhance performance (with an AUC improvement of up to 10%) in various scenarios of multilingual test data. Our results, which were validated on real-world data collected in three languages during ten disasters across seven countries, indicate the use of behavior-guided teacher models in MulTMR can increase attention to relevant indicators of serviceability characteristics. The application of the MulTMR framework through a streaming data analytics tool could reduce the cognitive load on personnel within social media teams of emergency services. Further, its application could inform how to leverage human behavior characteristics in creating automated models for social media analytics to design systems in other public service domains beyond emergency management.

## Introduction

Social media is instrumental in connecting the public with various organizations, such as governments, non-profits, and for-profit companies (Albanna, Alalwan, and Al-Emran 2022). In the case of for-profit companies, there has been a growing recognition of the value of providing customer service through social media. These companies often respond promptly to social media inquiries from both current and potential customers. Likewise, recent research indicates the public expects timely responses to their social media queries directed at governments and non-profit organizations (DiCarlo and Berglund 2020; Dahal, Idris, and Bravo 2021; Knox 2023).

| | Event | Serviceability | Message |
|---|---|---|---|
| M1 | Turkey-Syria Earthquake 2023 | serviceable (help request) | @SERVICE Hayrullah mahallesi 16. sokak'taki Ferhat apartmanında acil yardıma ihtiyaç var! **[EN]**@*SERVICE Urgent help needed at Ferhat apartment in 16th street, Hayrullah neighborhood!* |
| M2 | Hurricane Sandy 2012 | serviceable (information request) | @SERVICE how I can volunteer to help clean up after the hurricane? |
| M3 | Catalonia Fires 2019 | non-serviceable (gratitude and complaining) | @SERVICE Realizáis un gran trabajo y no os pagan lo suficiente por ello, de verdad muchas gracias **[EN]**@*SERVICE You guys do a great job, and you don't get paid enough for it, really thank you so much* |

Table 1: Examples of multilingual messages with varied serviceability characteristics that were directed at emergency services' accounts on a social media platform. (*Note: messages were paraphrased for anonymity.*)

To meet these expectations of the public, social media accounts of emergency services and non-profit organizations in particular face significant challenges. During disasters, the public posts an enormous number of messages on social media at a high velocity, leading to information overload for emergency services that have limited human resources (Kaufhold et al. 2020). Further, the value of these messages to services varies greatly, ranging from specific requests for information or resources and unsolicited offers of help to unsubstantiated rumors, concerns, and prayers that may not be serviceable requests (Purohit et al. 2018). Consequently, there is an urgent need for enhancing automated

[*]Part of this research was conducted during author's short-term scholar visit at George Mason University.

social media analytics systems for communication departments of public services to quickly prioritize messages that require a timely response. Moreover, there is a limited research on helping emergency services, especially in regions with low-resource languages, or multilingual, non-English-speaking populations on social media during disasters.

Table 1 demonstrates examples of various messages addressed to emergency services in different regions, cultures, and languages during disaster events. M1 is a prototypical serviceable message containing a concrete help request (informing the address where people need rescue). M2 is also serviceable that has a request for relevant information (asking how a user can become a volunteer). Finally, M3 is not a serviceable request from the perspective of operational response, but a message expressing gratitude and complaining. Capturing these various types of human behavior, along with understanding their prevalence in the content across diverse languages and regions, makes the task of automatically detecting and prioritizing a serviceable help request on social media challenging.

Our paper investigates the following research questions:

- RQ1. How can we train a classification and ranking model for a social media platform to process multilingual serviceable requests while learning different types of human behaviors when seeking help on social media?
- RQ2. To what extent does the performance improvement of the proposed framework depend on the types of behavior-guided models used?
- RQ3. Are there any differences in attention on various parts of a request content resulting from behavioral fine-tuning, to analyze the model's understanding of relevant human behaviors in multilingual requests?

To address these questions, our framework relies on the popular knowledge distillation process (Hinton, Vinyals, and Dean 2015) for designing a computational framework called *Multiple Teachers Model for detecting and Ranking (MulTMR)*, which can detect and prioritize multilingual serviceable help requests on social media for public services. This process aims to transfer knowledge from one or more complex models (like a *teacher*) to a simpler model (like a *student*) for a task, to train it to mimic the teacher models.

In the case of social media analytics for public services, involving cross-domain scenarios (e.g., new disaster event, language, region), target task-related teachers may encounter erroneous model decisions due to the challenges of limited understanding and access to the required knowledge from the data of source domain. In these dynamic and multifaceted situations, the model may struggle to adapt and make accurate decisions, as it grapples with the nuances and complexities of diverse contexts. However, the incorporation of behavior-guided teachers can play a pivotal role in mitigating these challenges because some patterns of users' behavior could be similar even during different types of languages and domains of disaster events, e.g., patterns for information needs, help-seeking intent, questioning. By harnessing behavior-guided teachers, we can positively influence the distribution of attention weights within the model, which helps to ensure the model's decision-making process becomes more robust and adaptable, even in complex and diverse linguistic and domain contexts. Thus, developing an architecture for synergy between task-related and behavior-guided teacher types could improve model performance. It creatively leverages behavior-guided teacher models in the knowledge distillation process for achieving higher performance on a task.

We utilize pre-trained language models that have been fine-tuned to identify sarcasm behavior and questioning behavior, which allows for more understanding of diverse user behaviors in help-seeking messages. The automated decision-making of the MulTMR is analyzed by comparing the distribution of attention weight maps within the textual messages. This novel framework enables the creation of an efficient classification and ranking system of multilingual serviceable help requests that utilizes multiple teachers with a focus on teaching behavior characteristics. The resulting framework demonstrates a high level of performance to capture different human behaviors in help-seeking and could be applied to build social media analytics systems for public services across languages, regions, and application domains.

## Related Work

In this section, we discuss studies that have been conducted on filtering and ranking serviceable help requests on social media. We will also provide an overview of related literature on multilingual text classification methods for disaster-related social media messages and the Teacher-Student modeling approach.

### Social Media Requests

The literature offers insights into modeling requesting behavior or information-seeking intent across various domains, such as forums, email communication, and social media platforms. Researchers have identified request behavior in online forums across diverse contexts, such as urgency, informational intent, and social support. Furthermore, social media has emerged as a widely used channel for seeking help when individuals face challenging situations, such as health problems (Gupta, Khan, and Kumar 2022; Khan and Loh 2022), mental disorders (Pretorius et al. 2020), and public health emergencies (Luo et al. 2020; Li et al. 2021a).

Similarly, during disasters also, social media has become a popular platform for users to seek help (Purohit et al. 2013; Nishikawa et al. 2018; Zade et al. 2018; Cheng, Liu, and Li 2020; DiCarlo and Berglund 2020). Whether it is for rescue, supplies, or critical information, social media accounts of public services are often the first point of contact for those in need. Unlike other online service scenarios, time-critical messages during disasters require immediate attention and need to be directed to the intended target, such as rescue teams, for timely offline responses. As a result, special strategies have been developed to ensure that serviceable messages requesting help receive the necessary attention (Purohit et al. 2018; Song and Fujishiro 2019; Purohit, Castillo, and Pandey 2020; Imran et al. 2020).

Researchers have studied the factors that influence the spread and response of requests on social media. They have

focused on two categories of features: content characteristics and creator characteristics. Relevant features, such as content type, emotional tone, proximity, depth of self-disclosure, and social capital of help seekers, have been explored to determine how they affect the popularity and effectiveness of messages that contain requests (Li et al. 2021b; Li, Bahursettiwar, and Kogan 2021; He et al. 2022). Studies on characterizing various types of user behavior when posting messages to seek help have also been conducted through theory-driven approaches. For example, researchers have explored how theories such as the negativity bias theory (Rozin and Royzman 2001) can be applied to help-seeking scenarios (Lifang et al. 2020).

### Classification of Multilingual Messages on Social Media During Disasters

In the context of disaster management, the ability to gather relevant and timely information from social media platforms is crucial (Purohit and Peterson 2020). However, this task becomes significantly more challenging when dealing with messages in multiple languages (Vitiugin and Castillo 2019; Lorini et al. 2021; Salemi, Senarath, and Purohit 2023).

To address this issue, recent studies have proposed approaches for retrieving and classifying information from disaster-related social media. These methods leverage deep-learning models with multilingual embeddings (Lorini et al. 2019; Kayi et al. 2020; Salemi, Senarath, and Purohit 2023) and large language models (LLMs), with the latter approach involving fine-tuning LLMs for specific tasks and domains (Liu et al. 2021; Sánchez et al. 2022). Another approach to utilizing LLMs for the purpose of classifying disaster-related messages involves employing a Teacher-Student training method within a cross-lingual transfer scenario to fine-tune models (Krishnan et al. 2022). In addition, researchers have proposed the combination of the versatility of graph neural networks, applied to a corpus, with the power of transformer-based LLMs, applied to examples through cross-attention (Ghosh, Maji, and Desarkar 2022). Another approach for emerging disaster-related social media involves the use of a multimodal neural network that effectively incorporates both textual and visual data (Koshy and Elango 2023), and specific information, such as hydrological (de Bruijn et al. 2020) data. By utilizing these techniques, emergency management personnel can more effectively analyze and interpret multilingual social media content during disasters, ultimately improving their ability to respond to and mitigate the adverse impact of these events (Vitiugin and Castillo 2022).

### Knowledge Distillation and Teacher-Student Model

The Teacher-Student model is a knowledge distillation approach (Hinton, Vinyals, and Dean 2015) that aims to transfer knowledge from a complex model ($Teacher$) to a simpler model ($Student$) and has been utilized for various tasks such as reducing the dimension of word embeddings (Shin, Yang, and Choi 2019), self-knowledge distillation (Hsieh et al. 2023), or contrastive learning (Chen et al. 2020). However, the knowledge learned from a single teacher may be limited and biased, which can result in a low-quality student model. To address this, a multi-teacher knowledge distillation framework has been proposed for pre-trained language model compression, enabling the training of high-quality student models from multiple teachers LLMs (Wu, Wu, and Huang 2021). Recently, a multilingual knowledge distillation approach has been proposed that transfers knowledge from high-performance monolingual models to a multilingual model using a Teacher-Student approach, which enables the model to learn from multiple monolingual models simultaneously, resulting in improved performance (Yang et al. 2022). Furthermore, the teacher models need not be limited to LLMs, as task-specific models can also be used to transfer specific behavioral knowledge to the student model (Kim and Hassan 2020).

Building upon the insights derived from the previous two approaches, we introduced the multiple task-/behavior-guided teachers model for classifying and ranking serviceable requests for help. Through the combination of these two types of teacher models, we aim to significantly reduce uncertainty of the student model, particularly in zero-shot scenarios. This becomes specifically crucial when dealing with test data in a different language that substantially differs from the training and validation datasets. Our approach seeks to enhance the model's robustness and adaptability for social media analytics systems in the face of such challenging cross-lingual variations, by creatively leveraging behavioral characteristics of the user-generated content.

## Method

This section introduces the framework of the *Multiple Teachers Model for detecting and Ranking (MulTMR)* and describes how it can be used for detection and prioritization of multilingual serviceable help requests during disasters. First, we present MulTMR framework for collaborative teaching of the student model. Second, we describe the method for behavioral fine-tuning of pre-trained multilingual LLMs using question type and sarcasm classification tasks to learn relevant user behaviors for detecting serviceable help requests.

### *MulTMR*: Multiple Teachers Model for Detecting and Ranking

Our framework is inspired from the task-related language model distillation process (Kim and Hassan 2020) using a diverse set of multiple teachers (Wu, Wu, and Huang 2021). Its architecture presented in Figure 1 has two loss functions for knowledge distillation: multi-teacher hidden loss and multi-teacher distillation loss.

The multi-teacher hidden loss transfers knowledge between hidden states of multiple teachers. Suppose there are $N$ teacher models, and each of them has $T$ Transformer layers. They collaboratively teach a student model with $T$ layers, and each $j$-th layer in the student model corresponds to $j$-th layer in a teacher model [1]. Denote the hidden states out-

[1]We maintain the same number of layers in the student model as the original teacher model – 24 layers for XML-RoBERTa (Conneau et al. 2019) and 12 layers for BERT (Devlin et al. 2018).

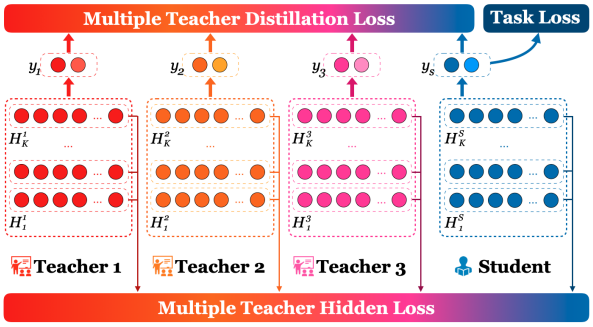



Figure 1: The overall architecture of MulTMR.

put by the $j$-th layer of the student model as $H_j^s$, and the corresponding hidden states output by the $j$-th layer of the $i$-th teacher model as $H_j^i$. We apply the mean squared error (MSE) to the hidden states of corresponding layers in the student and teacher models to encourage the student model to have similar functions with teacher models (Sun et al. 2019). The multi-teacher hidden loss $L_{hid}$ is formulated as:

$$\boldsymbol{L}_{hid} = \sum_{i=1}^{N}\sum_{j=1}^{T} MSE(H_j^S H_j^i W) \quad (1)$$

where $W$ is a set of hyper-parameters.

The multi-teacher distillation loss aims to transfer the knowledge in the soft labels output by multiple teachers to the student. The predictions of different teachers on the same sample may have different correctness and confidence. Since in task-related knowledge distillation the labels of training samples are available, we used a distillation loss weighting method to assign different weights to different teachers by grid search. The multi-teacher distillation loss $L_{dis}$ is formulated as follows:

$$\boldsymbol{L}_{dis} = \sum_{i=1}^{N} CE(y_s/t, y_i/t) \quad (2)$$

where $CE(\cdot,\cdot)$ stands for the cross-entropy loss, $y_s$ and $y_i$ are predictions by student and teachers models respectively, and $t$ is the temperature coefficient.

Next, we incorporate gold labels $y$ to compute the task-related loss on the predictions of the student model: $L_{task} = CE(y, y_s)$. The final loss function $L$ for learning the student model is a summation of the multi-teacher hidden loss, multi-teacher distillation loss and the task-related loss, which is formulated as follows:

$$\boldsymbol{L} = \alpha\boldsymbol{L}_{task} + (1-\alpha)\boldsymbol{L}_{dis} + \beta\boldsymbol{L}_{task} + (1-\beta)\boldsymbol{L}_{hid} \quad (3)$$

where $\alpha$ and $\beta$ are hyperparameters.

## Behavioral Fine-Tuning of Pre-Trained Models

Behavior-guided or behavioral fine-tuning (Ruder 2021), through a number of related tasks (Aghajanyan et al. 2021), refers to the process of teaching a model relevant capability that are useful for performing well on a target task, which requires understanding diverse patterns of human behavior from languages (Founta et al. 2019; Zhang et al. 2023).

| | Behavior type | Example | Type |
|---|---|---|---|
| B1 | Imperative mood requests | *Prohibit rockets, firecrackers, and dangerous activities with fires during this drought!* | FN |
| B1 | Imperative mood requests | *Many of us have no choice… I have to drive 80 km to go to work.* | FP |
| B2 | Sarcastic questions | *Now if you want the Spanish army to come in and get your chestnuts out of the fire, right?* | FP |
| B2 | Sarcastic questions | *What does it say? Sorry, but I don't speak Catalan and I want to find out* | FN |
| B3 | Short question | *Is there no more fire?* | FN |
| B4 | Information requests in a form close to a complaint | *You say we are strong together, you prevent aid. You cannot rule alone. People die while waiting for instructions. There are voices coming from under the buildings, but you are passing by… Is this unity????* | FN |
| B5 | Contextual requests | *Gazi Mustafa Kemal street No:50/A Opposite Güneşli mosque Elbistan, Kahramanmaraş* | FN |

Table 2: Erroneous examples in preliminary analysis. (*Note: messages were paraphrased for anonymity.*)

This is accomplished by fine-tuning the model on tasks related to the target task. It is called "behavioral" fine-tuning because it emphasizes the acquisition of practical behaviors, as opposed to adaptive fine-tuning. Particularly, behavioral fine-tuning using annotated data has proven effective in teaching models about various linguistic features such as named entities (Broscheit 2020), paraphrasing (Arase and Tsujii 2019), syntax (Glavaš and Vulić 2021), answer sentence selection (Garg, Vu, and Moschitti 2020), and question answering (Khashabi et al. 2020). A recent study on fine-tuning a model on nearly 50 annotated datasets in a massively multitask environment yielded the observation that a comprehensive and varied selection of tasks is crucial for achieving optimal transfer performance (Aghajanyan et al. 2021).

Detection or classification of serviceable request messages posted in social media during disasters is a challenging task because of the extreme variety of content presented in text data for expressing diverse user behaviors. Further, ranking of multilingual messages for serviceable help requests is a more challenging task because of the increase in syntactic and semantic redundancies, i.e., a multilingual model should be more context-sensitive and consider differences presented in user-generated content in different languages (Vitiugin and Castillo 2019). Our approach is based on an intuition of detecting and ranking serviceable requests for help using a behavioral fine-tuning approach, i.e., use of models for detecting specific behavior of users in a disaster relevant to a region or culture, or the type of disaster.

At first, we fine-tuned multilingual transformer-based model (Multilingual BERT (Devlin et al. 2018) and XML-RoBERTa (Conneau et al. 2019)) for detecting serviceable

| Event (start-end month/day) | Service | Non-Service |
|---|---|---|
| English | | |
| Hurricane Sandy 2012 (10/27-11/07) | 30 | 30 |
| Oklahoma Tornado 2013 (05/20-06/10) | 28 | 24 |
| Louisiana Floods 2016 (08/14-09/29) | 19 | 37 |
| Alberta Floods 2013 (06/21-07/05) | 190 | 624 |
| Nepal Earthquake 2015 (04/15-05/15) | 40 | 198 |
| Hurricane Harvey 2017 (08/29-09/15) | 209 | 1323 |
| Spanish | | |
| Catalonia Fires 2019 (06/04-06/30) | 28 | 163 |
| Chile Earthquake 2014 (04/02-04/07) | 358 | 1197 |
| Gloria Storm 2020 (01/26-01/28) | 32 | 44 |
| Turkish | | |
| Turkey-Syria Earthquake 2023 (02/05-02/07) | 980 | 701 |
| Total number of messages | 1914 | 4341 |

Table 3: Summary of datasets: dates of disaster events and number of serviceable/non-serviceable messages.

requests (task-related model). Next, we conducted an error analysis of common mistakes made by this task-related model. Table 2 demonstrates examples of the detected errors. Hence, we decided to address these mistakes by using additional teacher models as behavior-guided models during the distillation step: question type classification for the B1, B2, and B3 mistake types, sarcasm classification for the B4 type. For the B5 mistake type, we used a named entity recognition model during the pre-processing step and changed all location names by *LOCATION* tag. Based on our findings, we fine-tuned the same pre-trained language model (Multilingual BERT and XML-RoBERTa) for the two behavioral tasks. Finally, we had 3 fine-tuned models with the same architecture, tokenizers, and number of output classes. We describe the implementation details in the next section.

## Experiment Setup

In this section, we first describe the datasets used for the experiment and behavioral fine-tuning, the baselines and model variations, and the details of final implementation.

### Data

The data from Twitter (now X.com) platform for serviceable requests across multiple disasters in English were presented in a recent study (Purohit et al. 2018), while messages posted during Chile earthquake 2014 in Spanish were presented by CrisisNLP (Imran, Mitra, and Castillo 2016).

We also collected additional data in Spanish and Turkish via Twitter API. All collected tweets were annotated by one human assessor with language proficiency in the target language. In addition, we asked two persons native in Spanish and two persons native in Turkish to annotate 100 random messages in each language to calculate assessors' agreement. The Krippendorff's alpha for Spanish was 0.84 and for Turkish – 0.82. The annotation task was to assign one of the two classes for determining whether a given tweet is serviceable or non-serviceable for a target (such as emergency services like @emergenciescat, @AFADBaskanlik, @houstonpolice, etc.), using the similar setup as provided in prior studies (Purohit et al. 2018). For uncertain annotated texts, authors consulted an emergency service practitioner. Before annotating datasets, we conducted a simple preprocessing step (replaced mentions and URLs by corresponding special tokens), to filter out all uninformative tweets (based on manual analysis of 300 random messages, more than 90% of them with length $\leq 4$ words after removing special tokens are uninformative). Table 3 presents the quantity of train and test instances for each category.

To fine-tune pre-trained LLMs for knowledge distillation using behavior-guided models, we used existing public datasets:

- Sarcasm and irony detection dataset – contains 99000 English Tweets, 33000 of which contain the hashtag #irony or #ironic and 33000 contain #sarcasm or #sarcastic (Ling and Klinger 2016). We modified the dataset to fine-tune the pre-trained model for a binary classification. All messages from classes "sarcasm", "irony" and "figurative" were annotated as $sarcasm$, while the last class $regular$ stayed unmodified.
- Question type classification dataset – contains 5500 questions in 6 coarse classes ("abbreviation", "entity", "description", "human", "location" and "numeric value") (Li and Roth 2002). Based on the definitions of question classes, we annotated "description" and "location" as $relevant$, while other classes were annotated as $non-relevant$.

### Schemes

To evaluate our proposed framework, we compared it to the commonly used [2] pre-trained multilingual LLMs, and built a neural baseline model that utilized LSTM with DistilmBERT embeddings as input features. Both Multilingual BERT and XLM-RoBERTa models were used to evaluate the performance of our MulTMR framework. The full list of proposed modeling schemes for evaluation is the following (* denotes our proposed models and others are the baselines):

- ***[LSTM + DistilmBERT]*** – method uses pre-trained DistilmBERT sentence embeddings [3], which are passed as input to a Long Short-Term Memory Network model;
- ***[BERT]*** – BERT multilingual base model (cased) [4] was fine-tuned on the dataset with 5 frozen layers;
- ***[XLM-RoBERTa]*** – XLM-RoBERTa (large-sized model) [5] was fine-tuned on the dataset with 20 frozen layers;
- ***[*MulTMR-BERT]*** – Multiple Teachers Model for detecting and Ranking based on fine-tuned multilingual BERT model;
- ***[*MulTMR-RoBERTa]*** – Multiple Teachers Model for detecting and Ranking based on fine-tuned multilingual RoBERTa model.

[2] Based on HuggingFace.com downloads statistics.
[3] https://www.sbert.net/docs/pretrained_models.html
[4] https://huggingface.co/bert-base-multilingual-cased
[5] https://huggingface.co/xlm-roberta-large

| Model Scheme | ACC | F1 | AUC |
|---|---|---|---|
| *LSTM+DistilmBERT* | 80.92±0.65 | 62.21±5.74 | 85.74±3.40 |
| *BERT* | 80.85±2.38 | 81.19±2.09 | 80.04±0.99 |
| *XLM-RoBERTa* | 82.69±1.35 | 83.04±1.09 | 82.69±1.39 |
| **MulTMR-BERT* | 88.59±1.87 | 88.76±1.76 | 88.04±1.50 |
| **MulTMR-RoBERTa* | **88.97±1.23** | **89.07±1.19** | **88.23±1.16** |

Table 4: 5-fold Cross Validation (CV) results of the binary classification task. The best performances are in bold. Models were trained on multilingual data: train (67%) – validation (13%) – test (20%). $*$ denotes the proposed models.

We utilize three metrics to assess the effectiveness of binary classification models for serviceable requests detection, which are accuracy (ACC), area under the Receiver Operating Characteristic curve (AUC), and weighted F-measure (F1), in alignment with previous studies.

To compare the various schemes in learning to rank task for serviceable request ranking, we utilized the *normalized Discounted Cumulative Gain* ($nDCG$) metric, which provides a more significant weight to the discrepancies in the top positions compared to those occurring farther down the ranking outputs. Specifically, for each event (query) $i$:

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

where

- $\pi_i(j)$ – position of the document $d^i_j$ in ranking list $\pi_i$;
- $G^{-1}_{msx,i}(k)$ – normalizing factor at position $k$;
- $y_{i,j}$ – label of the document $d^i_j$ in ranking list $\pi_i$.

We evaluated $nDCG$ for the top-5, top-10, and top-20 ranked messages.

## Model Implementation

We maintain the same number of layers in the student model as the original teacher model: 24 layers for XML-RoBERTa (Conneau et al. 2019) and 12 layers for BERT (Devlin et al. 2018). During fine-tuning, we used the same hyperparameters and number of frozen layers (detected for task-related fine-tuning by grid search).

For LLMs' fine-tuning, we used $0.5 * 10^{-5}$ learning rate and 10 epochs. The number of frozen layers for each model were detected by grid search. For knowledge distillation, we used $0.6 * 10^{-5}$ learning rate and 10 epochs. Based on the results of grid search, we used the next hyperparameter values: $\alpha = 0.6$, $\beta = 0.5$, $t = 2$. The models were trained on NVIDIA A100-SXM4 with 40Gb GPU RAM via Google Colab.

# Result Analysis and Discussion

We first discuss the performance results of the proposed MulTMR schemes against the baseline schemes for research question RQ1, followed by an in-depth analysis of behavior-guided teacher models for RQ2, and the analysis of model interpretability for RQ3. Finally, we describe the analysis of cross-lingual classification scenarios and present the results for the learning to rank task as well.

| | baseline | +questions | +sarcasm | all |
|---|---|---|---|---|
| ACC | 80.85 | 87.82 | 87.94 | 88.59 |
| *Compare*: | | | | |
| - baseline | | 0.00021 | 0.00017 | 0.00014 |
| - all | | 0.4244 | 0.5225 | – |
| F1 | 81.19 | 87.83 | 88.03 | 88.76 |
| *Compare*: | | | | |
| - baseline | | 0.00013 | 0.00010 | 0.00008 |
| - all | | 0.3191 | 0.4099 | – |
| AUC | 80.04 | 85.91 | 86.90 | 88.04 |
| *Compare*: | | | | |
| - baseline | | 0.00001 | 0.00000 | 0.00000 |
| - all | | 0.02878 | 0.4268 | – |

Table 5: The impact of Behavior-guided models in *MulTMR-BERT* model is studied for the model schemes with the different sets of teachers. The table shows $p$-value for each scheme's performance in comparison with the baseline (*BERT*) and all-teachers model (*MulTMR-BERT*).

## MulTMR Performance

Table 4 displays the performance evaluation of MulTMR-based models against other schemes, to address RQ1. It is evident that both proposed models, MulTMR-BERT and MulTMR-RoBERTa, exhibit superior performance compared to the baselines across all metrics. Notably, the MulTMR-RoBERTa model performs better in multilingual settings, which could be attributed to the larger size of the underlying RoBERTa model in terms of the number of layers and parameters, as compared to BERT model.

Furthermore, MulTMR emphasizes the importance of behavior-guided models by incorporating additional features for serviceable help requests detection tasks. In comparison with the LSTM+DistilmBERT model, our architecture demonstrates superior performance, with an improvement of approximately 3% in AUC on multilingual data. Additionally, there are significant enhancements in accuracy and F1, with improvements of 8% and 27%, respectively. The primary reason for the higher performance is attributed to the knowledge distillation method used in MulTMR, which allows combining of hidden states and soft labels from multiple teachers, resulting in more accurate attention weights to achieve effective behavioral fine-tuning.

After analyzing the errors made by MulTMR, we discovered that the messages most commonly classified as false negatives were those that:

- were related to volunteering, including both offers and requests;
- were long messages that contained multiple thoughts, such as greetings and requests for information, simultaneously;
- were related to donations.

Similarly, the types of messages most commonly classified as false positive were those that:

- were long messages that contained multiple thoughts, such as reports and unclear help requests, simultaneously;
- contained complaints about the work of emergency services.

Using additional behavior-guided teacher models could improve the performance by avoiding such errors.

Apart from the higher performance, our framework design enables the highlighting of behavior-related tokens (analysis in Figure 2), which enhances the understanding of the model's generalizability, as described in the previous section.

## Impact of Behavior-Guided Models

To evaluate the impact of different teachers on the performance of the MulTMR model for addressing RQ2, we began with a baseline model that did not include any behavior-guided teacher models. We then designed two models by incorporating one behavior-guided teacher each for knowledge distillation. We used ACC, F1, and AUC as performance metrics to compare different model schemes, and the complete results can be found in the Table 5.

The results indicated that adding a second teacher to the knowledge distillation pipeline led to a statistically significant improvement in all measures. Adding just one behavior-guided teacher enhanced the model performance by 6-7%. We also compared the significance of the third teacher relative to the two previous ones. Our findings suggest that incorporating the sarcasm-detection teacher model is more significant in the AUC measure, implying that the knowledge about sarcasm provided by MulTMR is more valuable than knowledge about question types. This might indicate the model's effective understanding of user behavior for what to filter out in detecting serviceable help requests. Despite this, the use of the third teacher still led to an improvement of around 1% in the final model performance.

Based on our analysis, we conclude that the use of different behavior-guided teacher models results in faster convergence of model training and improved performance in the task of serviceable help requests detection.

## Analysis of Behavior-Guided Modeling

In this part of our study for addressing RQ3, we examined the relationship between behavior-guided modeling and attention weight maps generated by the MulTMR. Our objective was to investigate how MulTMR attention weights were related to behavior-guided modeling by analyzing the message texts written in each language from the dataset.

We utilized the MulTMR-BERT model to retrieve the attention weight maps by applying the MulTMR (with behavior-guided teachers) and without (baseline) on the texts. MulTMR-BERT model was chosen because it is faster, uses less memory, while still has comparable performance to MulTMR-RoBERTa. Our findings, as depicted in the Figure 2, show the attention weights of the MulTMR model focus more on valuable details. Additionally, this result was consistent for all three languages, highlighting the importance of knowledge obtained from behavior-guided teachers.

## Cross-Lingual Performance

In addition to multilingual classification tasks, there are also cross-lingual classification settings where the languages in the training and testing data are different. To assess the proposed framework's cross-lingual capability, we utilize a "leave-one-language-out" setting, where we train and validate the MulTMR on the data of a pair of languages and test it on the data of third language (e.g., train on English (EN) and Spanish (ES) and test on Turkish (TR)). To ensure unbiased results, we shuffled the training and validation data instances. The complete findings of the cross-lingual classification are outlined in Table 6.

In comparison to the baselines, the MulTMR exhibits improved performance for both non-English languages. The testing on Turkish yields an AUC of up to 73%, while Spanish yields almost 70% AUC, and English yields 67% AUC. The MulTMR results in a 3.5% AUC improvement compared to task-related models, and the MulTMR-RoBERTa model shows an improvement of 5% compared to LSTM+DistilmBERT.

The substantial performance variation observed in different language pairs can be attributed to the distinct characteristics of messages in each language. To illustrate, messages in Turkish indicated context-specific requests frequently (e.g., address only), whereas messages in Spanish often incorporated complaints about service speed and quality. While our approach demonstrates commendable performance in scenarios where it benefits from a setup of multilingual data processing, these performance variances become notably pronounced when employing the "leave-one-language-out" setup. This underscores how the inherent dissimilarities in message content across languages exert a significant influence on model performance in complex cases.

## The Sensitivity to Algorithmic Parameters

We assess the sensitivity of the performance of our approach in the "leave-one-language-out" settings to three algorithmic parameters — teachers' weights in the final loss, number of training epochs, and distillation temperature.

**Teachers' weights.** Choosing the right weights for teacher models in the final MulTMR model loss is crucial for the overall performance. So, the goal is to find an appropriate range, where we can achieve the best performance for ACC, F1, and AUC. We cannot visualize the 4th dimension, but the weight of the third teacher (Question type) could be calculated as:

$$\boldsymbol{W}_{question} = 1 - (W_{target} + W_{sarcasm}) \qquad (5)$$

Figure 3 suggests that [0.4, 0.3, 0.3] is our appropriate range where we achieve maximum performance based on all metrics. Clearly, there is a need for a balanced solution that will provide maximal relevant information and insights from each teacher to the student model, which we plan for future work.

**Number of training epochs.** Figure 4(a) demonstrates that $epoch = 10$ provides the best performance when experimented in the range [5:20]. Although, we did not run the experiments with the higher number of epochs because of potentially ineffective resource consumption.

**Distillation temperature.** The temperature controls the discrepancy between two distributions and can effectively determine the difficulty level of the distillation task (Li et al. 2023). During our experiment, we use temperature value in

| Schemes | EN & ES → TR | | | EN & TR → ES | | | ES & TR → EN | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | F1 | AUC | ACC | F1 | AUC | ACC | F1 | AUC | ACC | F1 | AUC |
| *LSTM+DistilmBERT* | 60.07 | 54.61 | 63.34 | 77.14 | 65.38 | 66.65 | 77.12 | 58.89 | 65.21 | 71.44 | 59.63 | 65.07 |
| *BERT* | 62.06 | 61.25 | 62.89 | 76.40 | 75.16 | 63.92 | 75.31 | 74.52 | 63.76 | 71.26 | 70.31 | 63.52 |
| *XML-RoBERTa* | 70.25 | 70.30 | 70.29 | 79.60 | 77.58 | 65.53 | 79.50 | 77.11 | 64.23 | 76.45 | 75.00 | 66.68 |
| **MulTMR-BERT* | 67.44 | 67.05 | 66.53 | 77.88 | 76.66 | 67.26 | 77.68 | 76.27 | 67.27 | 74.33 | 73.32 | 67.02 |
| **MulTMR-RoBERTa* | **73.93** | **73.82** | **73.58** | **81.71** | **80.03** | **69.86** | **81.14** | **79.63** | **67.10** | **78.92** | **77.83** | **70.18** |

Table 6: 5-fold CV results of the binary classification task across languages ("leave-one-language-out" setting): test data contains messages in one language, while the training data contains messages in other languages. $*$ denotes the proposed models.

the [1:5] interval. Figure 4(b) demonstrates that the best performance with $T = 4$. It is important to mention that keeping a constant temperature is usually suboptimal, and use of more flexible approaches could significantly improve performance of the student model.

## Learning to Rank Serviceable Help Requests

Finally, we present a supervised learning approach for automatically ranking serviceable request messages using MulTMR framework. The objective of automatic ranking is to prioritize a list of messages based on their serviceability characteristics. We used the learning-to-rank methodology (Liu et al. 2009) to achieve this goal. The learning-to-rank method aims to learn a ranking model that can associate each query with a permutation of documents that matches the training data labels for graded relevance as closely as possible. The documents that are deemed more serviceable receive higher graded labels and are associated with higher (better) positions in the relevance ranking. While the proposed method can accommodate any relevance grade levels of serviceability, given the annotated data we had, we used binary levels in this experiment. It should be noted that our approach is applicable to any number of relevance grade levels. To accomplish this, we have employed the LambdaMART algorithm (Wu et al. 2010; Burges 2010; Qin et al. 2020), which relies on Gradient Boosted Decision Trees.

To train the ranking models, we utilized the data from all events except one, which was reserved for testing the model using the "leave-one-event-out" approach. We utilized sentence embeddings (generated by baselines and MulTMR) as input features for ranking model. We then obtained rankings for messages in each event through the 5-fold cross-validation setting.

First, we compare MulTMR and baselines with Social-EOC (Purohit et al. 2018) model, which was based on text features and human-annotated features of serviceability characteristics (explicit request/answerable question, correctly addressed, sufficiently detailed). For this evaluation, we used data in English only, as used in the prior study (6 events). Table 7 demonstrates performance results. Experiments show better performance of MulTMR-RoBERTa model over baselines, even when additional human-annotated features are used by Social-EOC model.

Table 8 compares the performance of different schemes in terms of $nDCG$ of the first 5 positions ($nDCG@5$), 10 positions ($nDCG@10$), and 20 positions ($nDCG@20$). The results prove that MulTMR performs better than baseline models. MulTMR-RoBERTa outperforms in $nDCG@5$ and

| scheme | nDCG@5 | nDCG@10 | nDCG@20 |
|---|---|---|---|
| *Social-EOC* | 59.33 | 69.17 | 68.00 |
| *BERT* | 85.28 | 84.93 | 87.66 |
| *XML-RoBERTa* | 94.89 | 92.58 | 87.67 |
| **MulTMR-BERT* | 93.55 | 91.85 | 90.69 |
| **MulTMR-RoBERTa* | **96.83** | **95.20** | **93.03** |

Table 7: Comparison of the average $nDCG@5$, $nDCG@10$, and $nDCG@20$. 5-fold CV results for events in English. $*$ denotes the proposed models.

| scheme | nDCG@5 | nDCG@10 | nDCG@20 |
|---|---|---|---|
| *DistilmBERT* | 57.32 | 50.51 | 44.73 |
| *BERT* | 92.94 | 95.94 | 93.34 |
| *XML-RoBERTa* | 89.71 | 90.80 | 86.97 |
| **MulTMR-BERT* | 96.14 | 96.34 | **94.12** |
| **MulTMR-RoBERTa* | **99.48** | **97.70** | 93.58 |

Table 8: Comparison of the average $nDCG@5$, $nDCG@10$, and $nDCG@20$. 5-fold CV results for events in English, Spanish, and Turkish. $*$ denotes the proposed models.

$nDCG@10$, while MulTMR-BERT shows the best results in $nDCG@20$. As expected, the MulTMR-based ranking models exhibit superior performance compared to the pre-trained DistilmBERT-based ranking model, by taking advantage of the knowledge of behavioral characteristics of help requests in messages.

## Conclusions and Future Work

This paper introduced a design of the behavior-guided knowledge distillation framework, *Multiple Teachers Model for detecting and Ranking (MulTMR)*, to detect and rank multilingual serviceable requests for help on social media during disasters. The core idea of MulTMR is to combine task-related and behavior-guided fine-tuned LLMs as teacher models for distilling knowledge to train a student model by optimizing hidden and distillation losses. The utilization of behavior-guided models helps to reduce uncertainty of results produced by a task-related teacher model alone. MulTMR pays close attention to important parts in the context and learns to give higher attention to the potential elements of behavioral characteristics in serviceable request messages for classification and ranking tasks. Experiments on the dataset of 10 events in three languages show that the proposed model outperforms several baselines in classification and ranking tasks. We presented extensive analyses to show the value of knowledge distillation with multiple teachers guided by human behavior characteristics, which

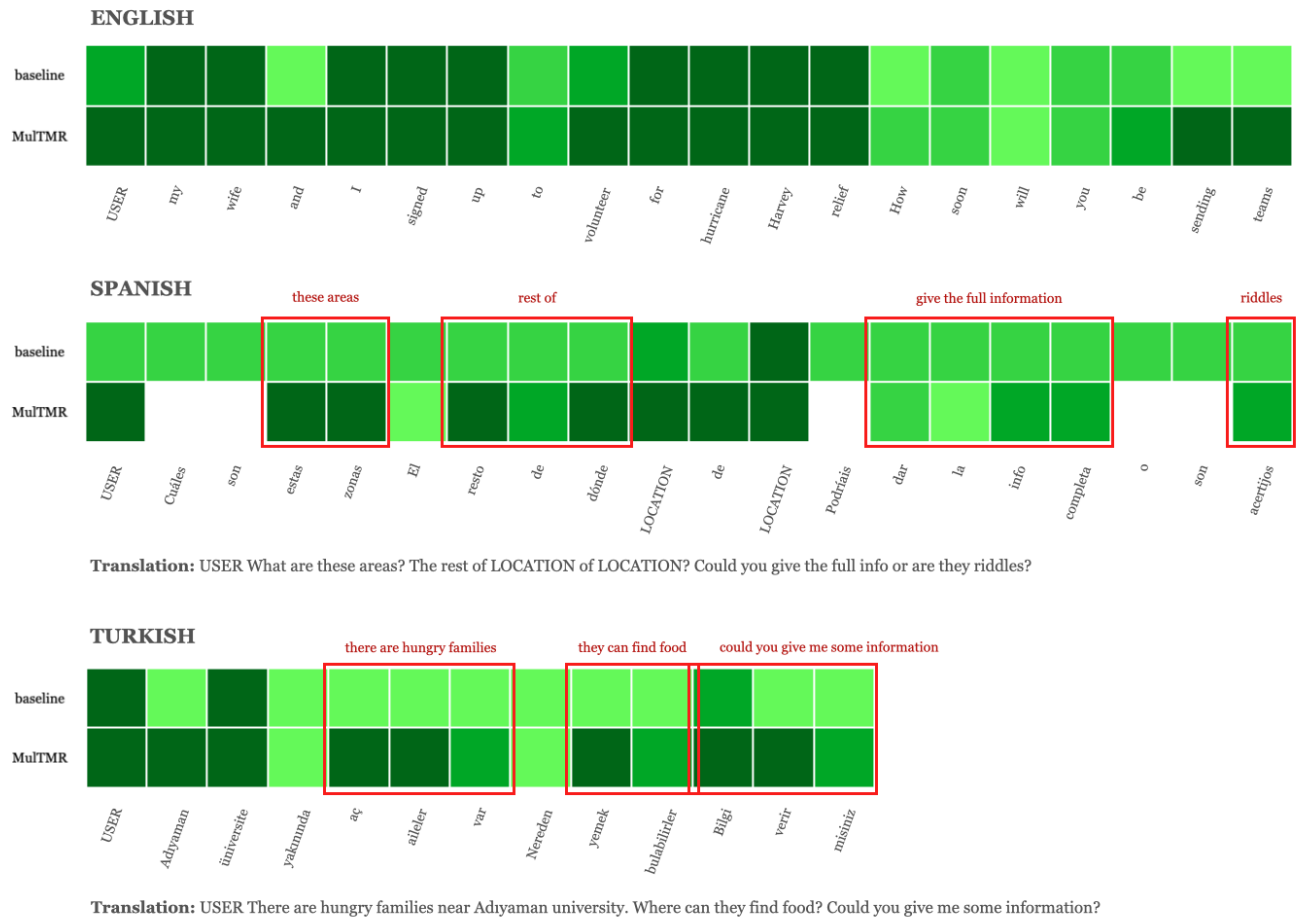



Figure 2: Attention weight maps of the texts in English, Spanish, and Turkish. The darker color indicates the higher weight.

can help adapt the model to different languages, event types, and tasks easily.

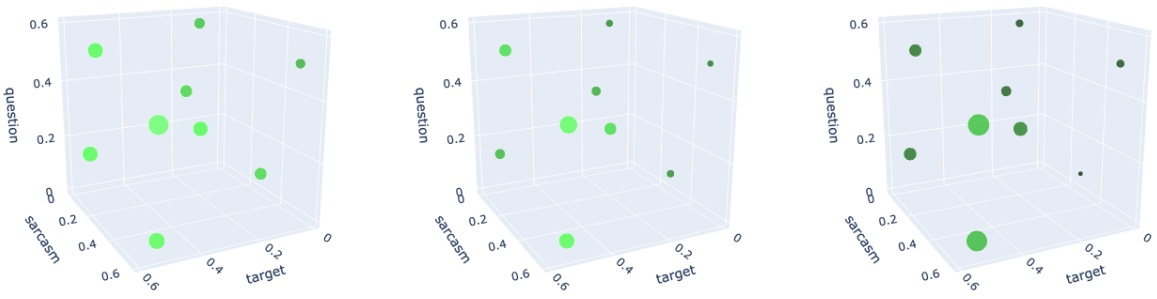
The MulTMR model holds promising potential for broader applications in social media-based services in different domains, with further research and enhancements to support specific organizations. Such service-related scenarios can span governmental agencies and businesses beyond emergency services, for instance, branch offices of post offices, public works, and more. By harnessing the capabilities of the MulTMR model, these entities can strengthen their customer relationships and potentially innovate new solutions, all grounded in the behavior analysis of the most serviceable requests by the public users. The versatility and adaptability of MulTMR with use of different behavior-related teachers beyond its initial scope could be one of the exploratory directions for future work.

There are certain limitations to our study that future work could address. First, the dataset we used for experimentation only contains serviceable help requests posted during four types of natural disasters. To improve the model's performance across various types of events, it would be valuable to extend this dataset to include other types of disasters. Second, we only used English, Spanish, and Turkish language messages in our experiments due to dataset limitations. Similarly, our experimental dataset includes messages posted during disasters in seven countries only. Future studies could expand on this by using data in more languages and from more regions of the world, and understand how different behavior-guided models could affect the MulTMR's performance. Third, our proposed model was tested using data collected over eleven years. During these years, the platforms and utilization of social media might have evolved, e.g., the length of Twitter (now X) messages was increased from 140 to 280 characters. Future research could study more recent or streaming data to train and test the model during new events and also, analyze the performance differences across the datasets from different periods.

**Reproducibility:** Datasets and code for the experiments described in this paper are available for research purposes at the public repository https://github.com/vitiugin/multmr.

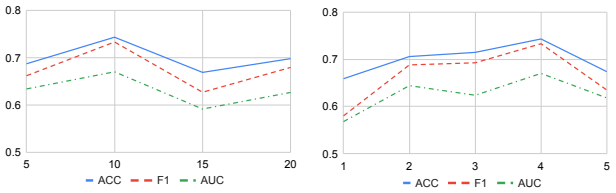
## Broader Impact and Ethics Statement

The primary motivation behind our research is driven by the expected broader impact to help various public services in the efficient detection and prioritization of help requests




(a) ACC (b) F1 (c) AUC

Figure 3: Teachers' weights in the final loss. The bigger size and lighter color of the point means the higher value of evaluation metrics.




(a) Training epochs (b) Distillation temperature

Figure 4: Hyperparameters sensibility.

from social media, especially emergency services with limited human resources during disasters. Timely identification of critical help requests is crucial for emergency services to respond effectively, even if such requests are multilingual. LLMs could help fill this gap and help develop useful frameworks. Although utilizing LLMs become very popular, they contain a potential risks such as bias amplification, misinformation dissemination, and privacy concerns. Therefore, this research employed the behavioral fine-tuning procedure to carefully leverage LLMs. While the first two risks are not closely related to the task of serviceable request classification and ranking, the framework presented in our work could minimize them. We developed a model that provides explainable results based on the attention weights assigned to the tokens of messages for the validity of model understanding. The presented framework used Teacher-Student knowledge distillation approach where teachers are behavior-guided models. The data used in this study was obtained from social media platforms for publicly accessible user-generated content. To address privacy concerns, we have only provided tweet IDs and all examples of messages cited in the paper were modified and anonymized. Practitioners should not directly apply our findings to any domain of ranking serviceable requests without testing the resulting model on a sample of their desired application data for performance validation. Our key takeaway is that combining task-related and behavior-guided models can have a significant impact on the classification and ranking of serviceable help request messages posted during disasters on social media.

## Acknowledgments

Purohit thanks the Office of Research Computing for providing computing resources and the Office of Research, Innovation, and Economic Impact Fund (ORIEI) for partially supporting this research through the grant # 215135 at George Mason University. Further, this work has been partially supported by: "la Caixa" Foundation (ID 100010434), under the agreement LCF/PR/PR16/51110009; the Ministry of Science and Innovation of Spain with project "COMCRISIS", reference code PID2019-109064GB-I00; and the EU-funded "SoBigData++" project, under Grant Agreement 871042. The authors also express their deep gratitude to Humanitarian Informatics Lab members who helped in experimental setup and provided valuable insights.

## References

Aghajanyan, A.; Gupta, A.; Shrivastava, A.; Chen, X.; Zettlemoyer, L.; and Gupta, S. 2021. Muppet: Massive Multi-task Representations with Pre-Finetuning. In *In Proc. of EMNLP*, 5799–5811.

Albanna, H.; Alalwan, A. A.; and Al-Emran, M. 2022. An integrated model for using social media applications in non-profit organizations. *International Journal of Information Management*, 63: 102452.

Arase, Y.; and Tsujii, J. 2019. Transfer Fine-Tuning: A BERT Case Study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5393–5404. Hong Kong, China: Association for Computational Linguistics.

Broscheit, S. 2020. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. *arXiv preprint arXiv:2003.05473*.

Burges, C. J. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581): 81.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual repre-

sentations. In *International conference on machine learning*, 1597–1607. PMLR.

Cheng, S.; Liu, L.; and Li, K. 2020. Explaining the Factors Influencing the Individuals' Continuance Intention to Seek Information on Weibo during Rainstorm Disasters. *International Journal of Environmental Research and Public Health*, 17(17): E6072–E6072.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Dahal, L.; Idris, M. S.; and Bravo, V. 2021. "It helped us, and it hurt us" The role of social media in shaping agency and action among youth in post-disaster Nepal. *Journal of Contingencies and Crisis Management*, 29(2): 217–225.

de Bruijn, J. A.; de Moel, H.; Weerts, A. H.; de Ruiter, M. C.; Basar, E.; Eilander, D.; and Aerts, J. C. 2020. Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. *Computers & Geosciences*, 140: 104485.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DiCarlo, M. F.; and Berglund, E. Z. 2020. Use of social media to seek and provide help in Hurricanes Florence and Michael. *Smart Cities*, 3(4): 1187–1218.

Founta, A. M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2019. A unified deep learning architecture for abuse detection. In *In Proc. of ACM WebSci*, 105–114.

Garg, S.; Vu, T.; and Moschitti, A. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7780–7788.

Ghosh, S.; Maji, S.; and Desarkar, M. S. 2022. GNoM: Graph Neural Network Enhanced Language Models for Disaster Related Multilingual Text Classification. In *In Proc. of ACM WebSci*, 55–65.

Glavaš, G.; and Vulić, I. 2021. Is Supervised Syntactic Parsing Beneficial for Language Understanding? An Empirical Investigation. arXiv:2008.06788.

Gupta, P.; Khan, A.; and Kumar, A. 2022. Social media use by patients in health care: a scoping review. *International Journal of Healthcare Management*, 15(2): 121–131.

He, C.; Deng, Y.; Yang, W.; and Li, B. 2022. " Help! Can You Hear Me?": Understanding How Help-Seeking Posts are Overwhelmed on Social Media during a Natural Disaster. *In Proc. of CSCW*, 6(CSCW2): 1–25.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-K.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. arXiv:2305.02301.

Imran, M.; Mitra, P.; and Castillo, C. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *In Proc. of LREC 2016*. Paris, France: European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Imran, M.; Ofli, F.; Caragea, D.; and Torralba, A. 2020. Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions.

Kaufhold, M.-A.; Rupp, N.; Reuter, C.; and Habdank, M. 2020. Mitigating information overload in social media during conflicts and crises: design and evaluation of a cross-platform alerting system. *Behaviour & Information Technology*, 39(3): 319–342.

Kayi, E. S.; Nan, L.; Qu, B.; Diab, M.; and McKeown, K. 2020. Detecting urgency status of crisis tweets: A transfer learning approach for low resource languages. In *In Proc. of COLING*, 4693–4703.

Khan, M. I.; and Loh, J. 2022. Benefits, challenges, and social impact of health care providers' adoption of social media. *Social Science Computer Review*, 40(6): 1631–1647.

Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; and Hajishirzi, H. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1896–1907. Online: Association for Computational Linguistics.

Kim, Y. J.; and Hassan, H. 2020. FastFormers: Highly Efficient Transformer Models for Natural Language Understanding. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, 149–158.

Knox, C. C. 2023. Local emergency management's use of social media during disasters: a case study of Hurricane Irma. *Disasters*, 47(2): 247–266.

Koshy, R.; and Elango, S. 2023. Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. *Neural Computing and Applications*, 35(2): 1607–1627.

Krishnan, J.; Anastasopoulos, A.; Purohit, H.; and Rangwala, H. 2022. Cross-lingual text classification of transliterated Hindi and Malayalam. In *2022 IEEE International Conference on Big Data (Big Data)*, 1850–1857. IEEE.

Li, L.; Aldosery, A.; Vitiugin, F.; Nathan, N.; Novillo-Ortiz, D.; Castillo, C.; and Kostkova, P. 2021a. The response of governments and public health agencies to COVID-19 pandemics on social media: a multi-country analysis of twitter discourse. *Frontiers in Public Health*, 1410.

Li, L.; Tian, J.; Zhang, Q.; and Zhou, J. 2021b. Influence of content and creator characteristics on sharing disaster-related information on social media. *Information & Management*, 58(5): 103489.

Li, X.; Bahursettiwar, A.; and Kogan, M. 2021. Hello? Is There Anybody in There? Analysis of Factors Promoting Response From Authoritative Sources in Crisis. *In Proc. of ACM CSCW*, 5(CSCW1): 1–21.

Li, X.; and Roth, D. 2002. Learning question classifiers. In *In Proc. of COLING*.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1504–1512.

Lifang, L.; Zhiqiang, W.; Zhang, Q.; and Hong, W. 2020. Effect of anger, anxiety, and sadness on the propagation scale of social media posts after natural disasters. *Information Processing & Management*, 57(6): 102313.

Ling, J.; and Klinger, R. 2016. An empirical, quantitative analysis of the differences between sarcasm and irony. In *In Proc. of ESWC*, 203–216. Springer.

Liu, J.; Singhal, T.; Blessing, L. T.; Wood, K. L.; and Lim, K. H. 2021. Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. In *In Proc. of ACM Hypertext*, 133–141.

Liu, T.-Y.; et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3): 225–331.

Lorini, V.; Castillo, C.; Dottori, F.; Kalas, M.; Nappo, D.; and Salamon, P. 2019. Integrating Social Media into a Pan-European Flood Awareness System: A Multilingual Approach. In *In Proc. of ISCRAM*, 646–659.

Lorini, V.; Castillo, C.; Peterson, S.; Rufolo, P.; Purohit, H.; Pajarito, D.; de Albuquerque, J. P.; and Buntain, C. 2021. Social media for emergency management: Opportunities and challenges at the intersection of research and practice. In *In Proc. of ISCRAM*, 772–777.

Luo, C.; Li, Y.; Chen, A.; and Tang, Y. 2020. What triggers online help-seeking retransmission during the COVID-19 period? Empirical evidence from Chinese social media. *Plos one*, 15(11): e0241465.

Nishikawa, S.; Tanaka, N.; Utsu, K.; and Uchida, O. 2018. Time trend analysis of "# Rescue" tweets during and after the 2017 northern Kyushu heavy rain disaster. In *In Proc. of IEEE ICT-DM*, 1–4. IEEE.

Pretorius, C.; McCashin, D.; Kavanagh, N.; and Coyle, D. 2020. Searching for mental health: a mixed-methods study of young people's online help-seeking. In *In Proc. of ACM CHI*, 1–13.

Purohit, H.; Castillo, C.; Diaz, F.; Sheth, A.; and Meier, P. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*, 19(1).

Purohit, H.; Castillo, C.; Imran, M.; and Pandey, R. 2018. Social-eoc: Serviceability model to rank social media requests for emergency operation centers. In *In Proc. of ASONAM*, 119–126. IEEE.

Purohit, H.; Castillo, C.; and Pandey, R. 2020. Ranking and grouping social media requests for emergency services using serviceability model. *Social Network Analysis and Mining*, 10: 1–17.

Purohit, H.; and Peterson, S. 2020. Social media mining for disaster management and community resilience. *Big data in emergency management: Exploitation techniques for social and mobile data*, 93–107.

Qin, Z.; Yan, L.; Zhuang, H.; Tay, Y.; Pasumarthi, R. K.; Wang, X.; Bendersky, M.; and Najork, M. 2020. Are neural rankers still outperformed by gradient boosted decision trees? In *International Conference on Learning Representations*.

Rozin, P.; and Royzman, E. B. 2001. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4): 296–320.

Ruder, S. 2021. Recent Advances in Language Model Fine-tuning. http://ruder.io/recent-advances-lm-fine-tuning.

Salemi, H.; Senarath, Y.; and Purohit, H. 2023. A Comparative Study of Pre-trained Language Models to Filter Informative Code-mixed Data on Social Media during Disasters. In *In Proc. of ISCRAM*, 920–932.

Sánchez, C.; Sarmiento, H.; Pérez, J.; Abeliuk, A.; and Poblete, B. 2022. Cross-Lingual and Cross-Domain Crisis Classification for Low-Resource Scenarios. *arXiv preprint arXiv:2209.02139*.

Shin, B.; Yang, H.; and Choi, J. D. 2019. The pupil has become the master: teacher-student model-based word embedding distillation with ensemble learning. In *In Proc. of IJCAI*, 3439–3445.

Song, C.; and Fujishiro, H. 2019. Toward the automatic detection of rescue-request tweets: Analyzing the features of data verified by the press. In *In Proc. of IEEE ICT-DM*, 1–4. IEEE.

Sun, S.; Cheng, Y.; Gan, Z.; and Liu, J. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.

Vitiugin, F.; and Castillo, C. 2019. Comparison of Social Media in English and Russian During Emergencies and Mass Convergence Events. In *In Proc. of ISCRAM*, 857–866.

Vitiugin, F.; and Castillo, C. 2022. Cross-Lingual Query-Based Summarization of Crisis-Related Social Media: An Abstractive Approach Using Transformers. In *In Proc. of ACM Hypertext*, 21–31.

Wu, C.; Wu, F.; and Huang, Y. 2021. One Teacher is Enough? Pre-trained Language Model Distillation from Multiple Teachers. In *In Proc. of ACL-IJCNLP*, 4408–4413.

Wu, Q.; Burges, C. J.; Svore, K. M.; and Gao, J. 2010. Adapting boosting for information retrieval measures. *Information Retrieval*, 13: 254–270.

Yang, Z.; Cui, Y.; Chen, Z.; and Wang, S. 2022. Cross-lingual text classification with multilingual distillation and zero-shot-aware training. *arXiv preprint arXiv:2202.13654*.

Zade, H.; Shah, K.; Rangarajan, V.; Kshirsagar, P.; Imran, M.; and Starbird, K. 2018. From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *In Proc. of ACM CSCW*, 2(CSCW): 1–18.

Zhang, D.; Li, W.; Niu, B.; and Wu, C. 2023. A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information. *Decision Support Systems*, 166: 113911.

## Paper Checklist

1. For most authors:
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No, we did not use any information on population-specific features, but we discuss possible artifacts in the data used.
   (e) Did you describe the limitations of your work? Yes, see the Conclusions and Future Work section
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see the Broader Impact and Ethics Statement
   (g) Did you discuss any potential misuse of your work? Yes, see the Broader Impact and Ethics Statement
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing:
   (a) Did you clearly state the assumptions underlying all theoretical results? NA, our study is empirical
   (b) Have you provided justifications for all theoretical results? NA, our study is empirical
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA, our study is empirical, but we discussed adaptability and explainability of the presented approach
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA, but we compared our approach with baseline and provide explanation of results
   (e) Did you address potential biases or limitations in your theoretical framework? NA, our study is empirical, but we provide limitations in Conclusions and Future Work section
   (f) Have you related your theoretical results to the existing literature in social science? NA, our study is empirical, but we discussed help-seeking behavior in Social Media Requests subsection in Related Work, see (Rozin and Royzman 2001) and (Lifang et al. 2020).
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA, but we discuss practical use in the Conclusions and Future Work and Broader Impact and Ethics Statement
3. Additionally, if you are including theoretical proofs:
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments:
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see the Reproducibility note and Model Implementation subsection
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see the Experiment Setup section
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, all experiments are fully reproducible with using random state in cross validation
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see the Model Implementation subsection
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see the Result Analysis and Discussion section
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we present insights on possible errors of misclassification and provide recommendations for practitioners in the Broader Impact and Ethics Statement
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**:
   (a) If your work uses existing assets, did you cite the creators? Yes, see the Data subsection
   (b) Did you mention the license of the assets? No, we will provide a license in the final version of github repository
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, as a URL in Reproducibility note
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see the Broader Impact and Ethics Statement
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see the Broader Impact and Ethics Statement
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets

FAIR? No, because we collected data via Twitter (now X.com) API during real-life events

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No, our dataset includes only Tweet IDs and labels

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**:

(a) Did you include the full text of instructions given to participants and screenshots? No, because data was annotated by one of the assessors who is professional in the task

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No, our dataset includes only IDs of public tweets

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No, all data was annotated voluntarily

(d) Did you discuss how data is stored, shared, and deidentified? Yes, see the Reproducibility note and Broader Impact and Ethics Statement