Social Network Analysis and Mining (2020) 10:58
https://doi.org/10.1007/s13278-020-00670-7

ORIGINAL ARTICLE

On detecting urgency in short crisis messages using minimal supervision and transfer learning

Mayank Kejriwal[1] · **Peilin Zhou**[1]

Received: 8 November 2019 / Revised: 20 June 2020 / Accepted: 23 June 2020 / Published online: 8 July 2020

Abstract
Humanitarian disasters have been on the rise in recent years due to the effects of climate change and socio-political situations such as the refugee crisis. Technology can be used to best mobilize resources such as food and water in the event of a natural disaster, by semi-automatically flagging tweets and short messages as indicating an urgent need. The problem is challenging not just because of the sparseness of data in the immediate aftermath of a disaster, but because of the varying characteristics of disasters in developing countries (making it difficult to train just one system) and the noise and quirks in social media. In this paper, we present a robust, low-supervision social media urgency system that adapts to arbitrary crises by leveraging both labeled and unlabeled data in an ensemble setting. The system is also able to adapt to new crises where an unlabeled background corpus may not be available yet by utilizing a simple and effective transfer learning methodology. Experimentally, our transfer learning and low-supervision approaches are found to outperform viable baselines with high significance on myriad disaster datasets.

Keywords Urgency detection · Social media · Machine learning · Twitter · Crisis informatics · Transfer learning

1 Introduction

The[1] United Nations Office for the Coordination of Human Affairs (OCHA) reported[2] that in 2018, more than 141 million people were in need of humanitarian assistance, with over 9 billion dollars of unmet requirements. Using technology to address this shortfall by assisting aid agencies and first responders to mobilize and send resources where they are needed the most is an important problem with the potential for widespread long-lasting social impact (Palen and Anderson 2016; Sakaki et al. 2010).

To achieve this goal, the problem of semi-automatic *urgency detection* needs to be solved, especially on short message streams like social media that support real-time news feeds and micro-updates from citizens on the ground. We define an urgent message in the crisis context as one that *expresses an actionable need that needs to be resolved in a short time frame*.

Urgency detection is related to the problem of detecting *relevant* or *informative* tweets from a stream of tweets, not all of which are pertinent to the crisis at hand. Urgency detection may be understood to be a very specific version of the relevance detection problem. Similar to the latter, urgency detection also falls in a class of information retrieval (IR) problems, which attempt to detect and rank relevant messages and documents. However, there is an added dimension to urgency detection, since (as defined above) an actionable need, possibly implied, must be expressed in the tweet that could potentially be resolved if dealt with in a time-sensitive manner. For example, a message such as 'Roof collapse in building on Main Street; multiple people trapped inside' may be deemed to be urgent; however, messages such as 'Roof collapse due to storm at midnight; all people successfully evacuated' and 'Avalanche in Nepal caused four deaths' are relevant and may assist in studying

✉ Mayank Kejriwal
kejriwal@isi.edu

[1] Marina del Rey, USA

[1] This article is an extended version of 'Low-supervision urgency detection and transfer in short crisis messages' (same authors) published in 2019 in the ASONAM conference. Unlike this article, that paper did not cover Research Question 2, which presents an algorithm for, and empirically investigates, transfer learning techniques for urgency detection in the minimally supervised setting.

[2] https://www.unocha.org/sites/unocha/files/WHDT2018_web_final_spread.pdf

Table 1 Urgent and non-urgent examples from three real-world datasets that we describe further in Sect. 5

Dataset	Urgent sentences	Non-urgent sentences
Nepal	Anyone who speaks about Balochistan in provinces other than Punjab either ends up dead or missing	Today's earthquake data for Nepal
	EMERGENCY: 4 locals trapped in this rubble INSIDE PALTANGHAR	Wow. ndtv just showed the same Philippines earthquake picture and said it's from Kathmandu on TV.
Macedonia	Some people are trapped in the marketplace need help.	the streets are filled with fecal and water no water
	We re trapped at the national commissioner s house the first floor s loaded with the kids have begun scared.	I'm about to walk with bicite but the rain that fell before s been blocking the roads that the channels are from the time of the rock.
Kerala	8 people no food survivin on dry cornflakes for the last 3 days east kadungalloor two families.	I'm from kerala and the situation here is very very bad, thousands have lost.
	At least 324 people have been killed in flooding and landslides in the indian state of while more than 200000	there has been floods in kerala india, more than 70 have lost their lives may 'Make it easy for all.'

the disaster further (or even mobilizing long-term response) but are not particularly urgent, either because it does not require immediate action or because the damage has already occurred. Informativeness as a broad problem has undergone some study (Olteanu et al. 2014) (see also Sect. 2), but to our knowledge, urgency as a specific IR area has not received the same kind of special attention despite its utility to first responders in times of crisis.

Put intuitively, solutions to the urgency detection problem can be framed in terms of probabilistic binary classification, a common machine learning paradigm involving other related tasks like sentiment analysis (Pang et al. 2008). Although urgency detection has some similarity with sentiment analysis (both are subjective to a degree, since annotators can, and do, sometimes disagree), the core problem is different, since the goal is to flag messages that *express urgency*, which is almost always a negative or panic-ridden emotion. However, it can be difficult to distinguish urgency-related tweets from just negative tweets. We provide an illustrative set of real-world examples[3] in Table 1.

In this paper, we present practical approaches for crisis-specific minimally supervised urgency detection on short message streams such as Twitter. The presented approaches cover two scenarios that often emerge in the real world. In the first scenario, a small amount (a few hundred messages) of training data labeled as urgent or non-urgent is available, along with a copious 'unlabeled' background corpus. In the second scenario, similar data are available for a 'source' domain but not for the target domain (expressing a 'new crisis') for which the urgency detection needs to be deployed. In other words, as messages are streaming in for this new domain, investigators label a few samples, but cannot rely on the availability of a background corpus since urgency needs to be tagged in real time before the crisis has fully subsided. To accomplish this challenging goal, our approach relies on a simple and robust transfer learning methodology (Pan and Yang 2010). Experimental results on three real-world datasets and several performance metrics validate our methods. To the best of our knowledge, this is the first such paper investigating the problem of urgency detection in social media, both algorithmically and empirically, for arbitrary disasters in low-supervision and transfer learning settings.

The rest of this paper is structured as follows. Section 2 describes some related work, Sect. 3 specifies our two research questions, and Sect. 4 describes our approaches in support of answering those questions. Section 5 covers the experiments, and Sect. 6 concludes the paper.

2 Related work

2.1 Crisis informatics and situational awareness

Crisis informatics is emerging as an important field for both data scientists and policy analysts. A good introduction to the field was provided in a recent Science policy forum article (Palen and Anderson 2016). The field draws on interdisciplinary strands of research, especially with respect to collecting, processing and analyzing real-world data. Particularly, social media platforms like Twitter have emerged as important channels ('social sensors' Sakaki et al. 2010) for *situational awareness* in support of crisis informatics. Although situational awareness is a broad notion extending beyond crisis informatics (e.g., military situational awareness), urgency detection is a special kind of situational awareness that tends to arise mainly in the crisis domain. A direct application is to help first responders and aid agencies assess needs in crisis-stricken areas and mobilize resources effectively (i.e., where needs are most urgent).

[3] A description of the datasets, as well as a link to the trained model itself, will be provided in Sect. 5.1.

While the initial primary focus of situational awareness and sensing systems was on earthquakes (Avvenuti et al. 2014; Crooks et al. 2013), the focus has diversified in recent years to disasters as diverse as floods, fire, and hurricanes (Arthur et al. 2017; Vieweg et al. 2010). We note that Twitter is by far the most monitored social media platform during crises (Simon et al. 2015) due to the availability of the published data and its real-time nature. Increasingly sophisticated approaches have been presented for data collection, including dynamic lexicons (Olteanu et al. 2014) and analysis tools like TweetTracker (Kumar et al. 2011).

In the last few years, and even just the last few weeks (in the wake of the COVID-19) crisis, a number of important works in network science have addressed crises. We only cite a few recent papers by way of reference. Recently, (Purohit et al. 2020) have described a method to rank and group social media requests for emergency services, a work that is particularly relevant since the outbreak of COVID-19. Recent work in opinion mining (e.g., see Keyvanpour et al. 2020), especially using lexicons and machine learning in social media, is also relevant to our work. Another extremely relevant work is a recent article that described a lightweight and multilingual framework for crisis information extraction from Twitter data (Interdonato et al. 2019). The research presented in that paper, though not resolving the problem of detecting urgent tweets, is compatible with our own work since it presents a relatively unsupervised and lightweight paradigm, and uses similar metrics. Other papers have tried to look at specific crises, e.g., the work by (Ladner et al. 2019) in analyzing tweets to determine the activeness of the Syrian refugee crisis. Another article has tried to do disaster damage assessment from Twitter data using statistical features and 'informative words,' not dissimilar to our own lexicon-based approach (Madichetty and Sridevi 2019). A last example is the work in (Klein et al. 2012), which describes a project called SABESS that uses social network analysis for identifying reliable tweets and apply content analysis in order to summarize important 'emergency facts.' These six examples are among a sample of several pieces of work that have tried to use social media productively in helping to analyze or provide actionable intelligence during a crisis situation attesting to the ongoing importance of the problem.

2.2 Crowdsourcing and existing crisis informatics platforms

NLP methods have been widely used in extracting situational awareness from Twitter, e.g., see the work by (Verma et al. 2011). Another important line of work is in analyzing events other than natural disasters (such as mass convergence and disruption events), but still relevant to crisis informatics. For example, Stabird et al. presented a collaborative filtering system for identifying on-the-ground 'Twitterers' during mass disruptions (Starbird et al. 2012). Similar techniques could be employed to supplement the work in this paper.

In a similar vein, the CrisisTracker system (Rogstadius et al. 2013) is another example of a system that uses crowdsourced social media curation for disaster awareness. The system does not specifically address the urgency detection problem, however. AIDR is a system that is more closely aligned with the goal of using AI for better disaster response (Imran et al. 2014), but its goal is to classify messages into a set of user-defined categories of information such as 'needs' and 'damages.' In contrast, we consider needs at a higher-level of classification; namely, is it urgent or non-urgent? The outputs of AIDR are compatible with our own since both systems provide actionable information to first responders.

Another important crowdsourcing tool that has been especially useful in working with SMS messages is Ushahidi, a project that was a grassroots effort that started in Kenya and that was used initially to encourage Kenyans to report incidents (especially, acts of violence) that they have witnessed. The website was very successful, and the model has since been replicated in other countries. Just like the other systems considered in this section, we believe Ushahidi's goals and technology are compatible with the capabilities presented herein.

More generally, projects like CrisisLex, Crisis Computing[4] and EPIC (Empowering the Public with Information in Crisis) have emerged as major efforts in the crisis informatics space due to two reasons: First, the abundance and fine granularity of social media data implies that mining such data during crises can lead to robust, real-time responses; second, the recognition that any technology that is thus developed must also address the inherent challenges (including problems of noise, scale and irrelevance) in working with such datasets. CrisisLex provides a repository of crisis-related social media data and tools, including collections of crisis data and lexicons of crisis terms (Olteanu et al. 2014). It also includes tools to help users create their own collections and lexicons. In contrast, Project EPIC, launched in 2009 and supported by a US National Science Foundation grant, is a multi-disciplinary effort involving several universities and languages with the goal of utilizing behavioral and technical knowledge of computer-mediated communication for better crisis study and emergency response. Since its founding, Project EPIC has led to several advances in the crisis informatics space; see, for example, (Barrenechea et al. 2015; Palen et al. 2015; Kogan et al. 2015; Anderson et al. 2013; Soden et al. 2014).

The work presented in this article is intended to be compatible with these efforts, although we are addressing a

[4] https://crisiscomputing.qcri.org/

specific problem that was not addressed by any of the works cited above. We have released our model openly, and potentially, this released model could be integrated into some of the platforms described above. Crowdsourcing could be used in lieu of (or even in addition to) the active learning framework presented as one of the solutions to the low supervision challenge described later in this article. It could also be used to provide more confidence in the annotations, since there is an inherent element of subjectivity when one is labeling a tweet as 'urgent.' Note that most labeling problems in machine learning involve some subjectivity, and inter-annotation agreement has been found to be a concern in some cases. Whether such concerns arise in the case of urgency detection is an unknown issue that does not fall within the scope of the presented work, but could be a valuable issue to address in future research.

2.3 Representation, transfer and deep learning in NLP and social media

Other lines of work relevant to this paper involve minimally supervised machine learning, representation learning and transfer learning. Concerning minimally supervised machine learning (ML), in general, ML techniques where there are few, and in the case of zero-shot learning (Palatucci et al. 2009; Romera-Paredes and Torr 2015), no observed instances for a label has been a popular research agenda for many years (Uszkoreit et al. 2009; Aggarwal and Zhai 2012). In addition to weak supervision approaches Aggarwal and Zhai (2012), both semi-supervised and active learning have also been studied in great depth, with surveys provided by (Zhu 2005; Settles 2010). However, to the best of our knowledge, a successful systems-level conjunction of various minimally supervised ML techniques has not been achieved for the task of short-text urgency detection. Such as empirical assessment is an important goal of this paper.

Due to the current renaissance of neural networks (Sahlgren 2005), *embedding* and *representation learning* methods have become more popular due to the advent of fast and effective models like skip-gram. Recent work has used such embeddings in numerous NLP and graph-theoretic applications (Collobert et al. 2011), including information extraction (Kejriwal and Szekely 2017), named entity recognition (Nadeau and Sekine 2007) and entity linking (Moro et al. 2014). The most well-known example is word2vec (for words) (Mikolov et al. 2013), followed by similar models like paragraph2vec (for multi-word text) and fasttext (Dai et al. 2015; Joulin et al. 2016), the last two being most relevant for the work in this paper. For a recent evaluation study on representation learning for text, including potential problems, we refer the reader to (Faruqui et al. 2016).

Finally, transfer learning is a central agenda in this paper; an excellent survey of dominant techniques may be found in (Pan and Yang 2010). More recent work on domain adaptation may be found in (Alam et al. 2018), with the work in (Pedrood and Purohit 2018) applied specifically to the disaster response problem. Pedrood and Purohit (2018) also applied transfer learning to the problem of mining help intent on Twitter. Other relevant work in crisis informatics, both in terms of defining 'actionable information' problems like urgency and need mining, and providing multimodal Twitter datasets from natural disasters, may be found in (He et al. 2017; Purohit et al. 2018) and Alam et al. (2018). Caragea et al., for example, present an approach to identifying informative messages in crises by using CNNs (Caragea et al. 2016). Other similar works include Burel et al. (2017) Burel and Alani (2018), Nguyen et al. (2016), (2017) and (Kersten et al. 2019). An important difference between the class of papers cited and our own work is that we are not seeking to detect *events* in crisis situations, but are instead trying to assign *urgency* scores to sub-events that are happening in the aftermath of a disaster. The two problems are related in that better accuracy on event detection (for which these deep learning systems could be used to great effect) would lead to better identification of urgent events. However, urgency detection is a difficult problem in and of itself, beyond the broader problem of isolating informative events related to the disaster from an incoming stream of messages. Some of the work above is multimodal (e.g., the paper by Nguyen et al. (2017)), which would be an interesting direction for future research for urgency detection (from images and videos, rather than just text).

An alternate way of looking at the problem is as an 'event detection' problem, e.g., in (Zheng et al. 2017) Zheng et al. study semi-supervised event-related tweet identification which also tries to identify the urgent tweets related to earthquakes and floods. These works are complementary to the minimally supervised, low-resource setting in this paper. Finally, we note that there has been some very recent work in few-shot models that use little to no training data and are similar to this paper in that regard Alam et al. (2018), Kruspe et al. (2019). However, there are significant differences from our own work. For example, while Alam et al. (2018) consider a specific disaster situation (flood risk assessment in a particular city in India), Kruspe et al. (2019) considers the earlier problem of detecting tweets that are relevant to the crisis itself, rather than the problem of assigning an urgency score to events that are detected. In general, we are not aware of a few-shot or minimally supervised technique that tackles urgency detection for the purposes of triage.

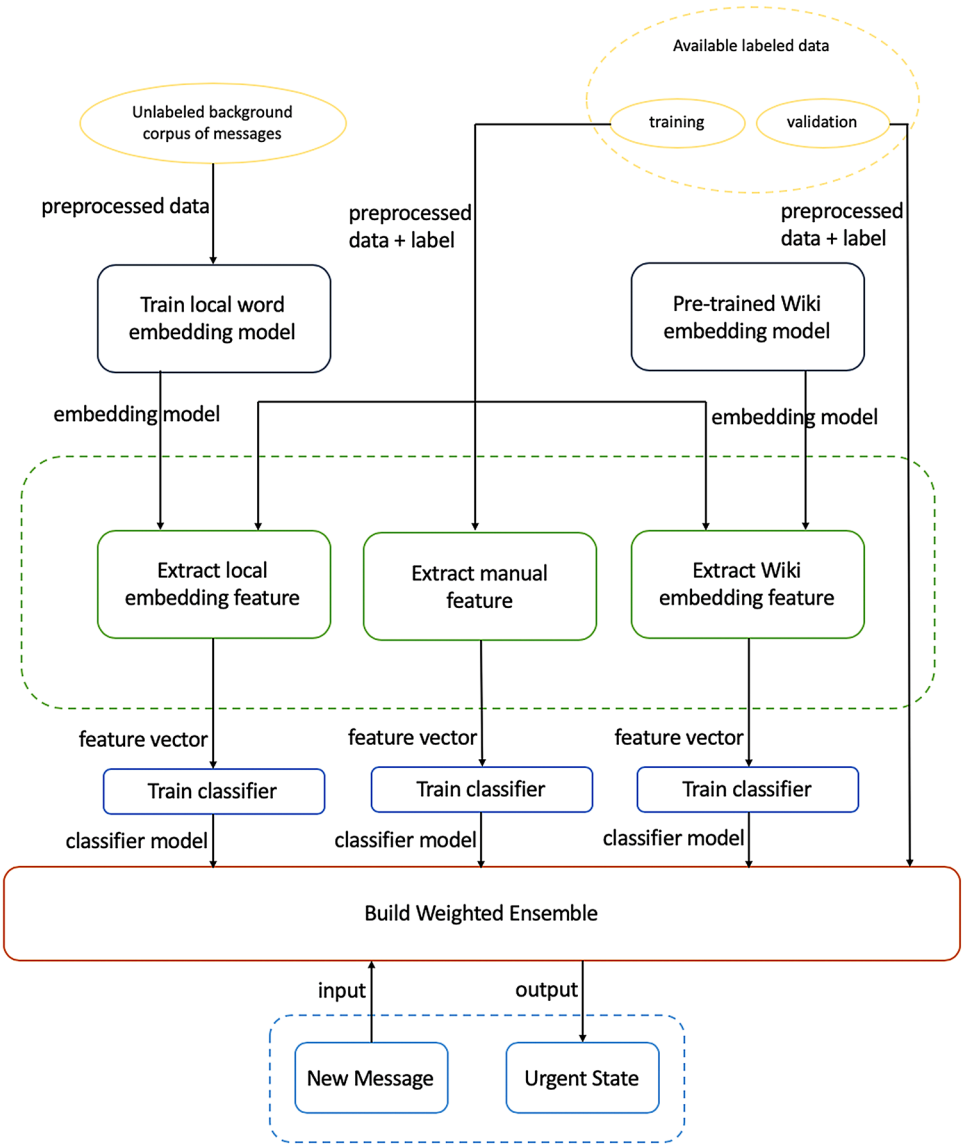

Fig. 1 Training workflow for urgency detection

3 Research questions

We briefly enumerate below the research questions under consideration in this paper. While the first question captures the classical low-supervision setting, the second question introduces an element of transfer learning.

1. **Low-supervision Training for Urgency Detection:** How do we build an urgency detection system for a specific crisis when given as training input both a small number of manually labeled tweets and a large number of unlabeled tweets (background corpus) for that crisis?
2. **Low-supervision Transfer Learning for Urgency Detection:** How do we build an urgency detection system for a specific crisis when given as training input a small number of manually labeled tweets for that crisis, as well as 'auxiliary' training input of (a small number of) manually labeled tweets and unlabeled background tweets from a *different* crisis?

Unlike the first scenario, the second scenario applies to a very short period (hours, or even minutes) after the crisis has struck; this is why a background corpus is not available (yet) for that crisis. Instead, only a few manually labeled messages that have been acquired till that point are available.

4 Approach

4.1 Low-supervision urgency detection

The approach for addressing the first research question is schematized in Fig. 1. The first step in the workflow involves data preprocessing of the corpus. We follow a standard set

of preprocessing steps. First, we apply a tokenizer to split the sentences into lists of words and delete words with special prefixes (including @ and RT, which are particularly prevalent in Twitter) and special suffixes. We also remove non-alphanumeric characters and convert the entire sentence to lowercase. Next, similar to traditional machine learning pipelines, we extract a set of manual features for expressing prior human knowledge about urgency detection. Our manual features are thus called because they are primarily keyword-based and binary, with keywords selected based on data exploration and domain knowledge. We consider ten such keywords, namely *hit, help, kill, injure, strand, miss, urgent, die, need, food*. If any of these keywords are present[5], the corresponding feature is set to 1. Note that these keywords are associated with situations that are generally urgent, like people who have been attacked or affected by a crisis and need urgent help, but some are noisier than others[6]. Additionally, we also utilize an eleventh feature that checks to see if any numeric digits are present in the dataset. The rationale behind this feature is that, in more urgent tweets, numbers are often present, e.g., '15 climbers are currently trapped on Everest due to the avalanche.'

In the Experiments section, we show that the manual features are not adequate for addressing low-supervision urgency detection. Besides, it is prudent to utilize the large number of unlabeled tweets (background corpus) if it serves a useful purpose in improving performance. To that end, we train a skip-gram based word embedding model based on the 'bag of tricks' model released by researchers from Facebook in a package called *fastText*Joulin et al. (2016). The reason behind using fastText, as opposed to alternate word embedding models like GloVe and word2vec Mikolov et al. (2013), is several-fold. First, fastText is very fast and easy to execute and is well maintained. Second, preliminary analyses showed that it does quite well on social media tasks and because of the bag of tricks methodology (that uses character and sub-word embeddings to gracefully deal with OOVs[7] and misspellings), it is able to generalize much better. Finally, fastText's APIs include a way to get sentence embeddings directly after training the word embedding model. By training fastText on the background corpus, we are able to train a robust embedding model. In both the training and test phases, we use this model to get feature vectors for our messages besides the 11-dimensional manual feature vector described earlier.

However, given that the background corpus might not be as extensive or representative as a 'general' corpus like Wikipedia, we try to smooth the feature space by also using a pre-trained embedding model trained over the English Wikipedia corpus and publicly available[8]. The vectors obtained from this model have 300 dimensions and were trained using skip gram with default parameters.

As Fig. 1 illustrates, we use all of these feature sets to build an ensemble by combining local embedding features, manual features and Wikipedia pre-trained word embedding features. The final score of the ensemble model is achieved by weighting the scores of the three linear regression models (one for each feature set), with weights adding to 1. The weights are set using a held-out validation set.

When the urgency of a new 'test' message needs to be determined, we preprocess the message, extract all three feature sets[9] and get the weighted score from the three regression models. If the score falls above a pre-determined threshold (again, determined through validation), then the message is flagged as urgent; otherwise, it is not.

4.2 Urgency detection using transfer learning

In this section, we describe our approach for 'urgency detection transfer' whereby a *source* dataset is given (similar to RQ1, where both an unlabeled background corpus and a small manually labeled training set are available) along with a *target* dataset (only a small manually labeled training set and no background corpus), representing the crisis under investigation. Our approach for urgency transfer is captured in Algorithm 1. Many of the steps are similar to those for RQ1, including preprocessing, but there are some important differences; hence, we use pseudocode to express the workflow more precisely. For example, while the Wiki embedding model remains the same as earlier, the manual features are obviously extracted over the target domain (since they do not require a background corpus) and importantly, the 'local' embedding model is now trained over the source domain corpus, since there is no target domain unlabeled background corpus available.

To 'sync' the source and target domains, we consider a simple, but empirically effective, approach. Rather than use just the labeled target domain data for training the three linear regression models, we combine the labeled training data from both the source and target domains, but the target training data are up-sampled to allow its properties to emerge more concretely in the training. The up-sampling margin is a parameter in Algorithm 1; in practice, a factor of 6 (meaning the target labeled dataset is up-sampled by 6x)

[5] Possibly as stems, for example, the word 'helping' would trigger the 'help' keyword feature, which would be consequently set to 1.

[6] For example, 'help' could be associated with a more trivial situation like someone needing help with their dog.

[7] Out of Vocabulary words.

[8] https://fasttext.cc/docs/en/pretrained-vectors.html

[9] In the case of the two trained embedding models, by getting the respective sentence embeddings for the test message

Table 2 Details on datasets used for experiments

Dataset	Unlabeled/ labeled messages	Urgent/non-urgent messages	Unique tokens	Avg. tokens/ message	Time range
Nepal	6,063/400	201/199	1,641	14	04/05/2015–05/06/2015
Macedonia	0/205	92/113	129	18	09/18/2018–09/21/2018
Kerala	92,046/400	125/275	19,393	15	08/17/2018–08/22/2018

has been found to work well. To maximize training dataset utility, we do not use a validation set for classifier weight optimization, but consider the average of all three classifiers as the final score.

Algorithm 1 Transfer Learning for Urgency Detection.

Input :

- Labeled dataset in target domain: D_t
- Labeled dataset in source domain: D_{sl}
- Unlabeled corpus in source domain: D_{su}
- Pre-trained Wikipedia Embedding Model: W_w
- Up-sampling parameter: u

Output :

- Classifier for Urgency Detection: $\mathcal{C}$

Method :

1. Train word embedding W_s on text in $D_{su} \cup D_{sl}$;
2. Up-sample D_t by factor u and 'mix' with D_{sl} to get expanded training set, $D_{train} : D_{tu} \cup D_s l$
3. Extract manual feature set F_m, source embedding feature set F_s (using W_s), and Wiki feature set F_w (using W_w) from each message in D_{train};
4. Train linear regression models C_s, C_m and C_w on F_s, F_m and F_w resp. to get classifier;
5. Return final classifier model $\mathcal{C} : avg\_score(C_s, C_m, C_w)$;

5 Experiments

5.1 Data

For evaluating the approaches laid out in Sect. 4, we consider three real-world datasets described in Table 2.

Two of the datasets (Nepal and Macedonia) were made available to us through the DARPA LORELEI program[10], under which this project is funded. The Nepal dataset comprises a collection of tweets collected in the aftermath of the 2015 Nepal earthquake (also called the Gorkha earthquake), while Macedonia was not an actual disaster but a realistic live-action simulation (of a disaster) conduced in Macedonia toward the end of 2018. Macedonia does not have much noise and is 'information-dense,' but small. As such, it provides a good test of the transfer learning abilities of the approach presented. Kerala describes tweets in the aftermath of the Kerala floods in South India in 2018 and is the largest dataset, with many relevant and irrelevant tweets. We note that these datasets were collected independently by an external participant in the program and made available to all performers in the program for research.

[10] https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

Table 3 Description for each baseline on research question 1

Baseline	Description
Local embedding (Local)	The features for a single linear classifier are sentence embeddings (with each pre-processed message treated as a 'sentence') trained using the 5-gram skip gram-based fastText model with vector dimensionality set to 20
Manual feature-based (Manual)	This baseline only considers the 11 manual features described earlier in Sect. 4
Wikipedia word embedding (Wiki)	This baseline only considers the linear classifier trained on the pre-trained Wikipedia embedding model
Local embedding and manual feature-based ensemble (Local-Manual)	This baseline combines *Local* and *Manual* by training two linear regression classifiers and weighting their probabilities to get the final result (using the validation set).
Local embedding and Wikipedia word embedding ensemble (Wiki-Local)	This baseline combines *Local* and *Wiki* using the same methodology as for *Local-Manual*.
Wikipedia word embedding and manual feature-based ensemble (Wiki-Manual)	This baseline combines *Manual* and *Wiki* using the same methodology as for *Local-Manual*.

Originally, all the raw messages for the datasets described in Table 2 were unlabeled, in that their urgency status was unknown. Since the Macedonia dataset only contains 205 messages and is a small but information-dense dataset, we labeled all messages in Macedonia as urgent or non-urgent (hence, there are no unlabeled messages in Macedonia as given in Table 2). For the two other Twitter-based datasets, we used active learning to compose a labeled set that would contain challenging examples. The basic process was to do data preprocessing as described in Sect. 4, followed by training the local fastText-based word embedding model on all messages in the corpus. Next, we randomly labeled 50 urgent and non-urgent tweets and fed them into a classifier. The classifier was applied on the rest of the unlabeled data to obtain 'ambiguous' examples (where the classifier's probability of the positive label was closest to 50%). We labeled another 100 samples this way and continued to re-train and apply the classifier for two more iterations till we obtained a total of 400 labeled points[11]. Note that the final labeled dataset may not be balanced in terms of urgent and non-urgent messages. Table 2 shows that Nepal is roughly balanced, while Kerala is imbalanced. We used stratified sampling therefore to split the labeled pool into training and testing datasets for evaluating the two research questions. We used 90% for training and 10% for testing. While the data cannot be made publicly available due to privacy concerns, we have released both the trained models and instructions for how to re-train the model on novel datasets as a Docker container[12].

5.2 Metrics

We consider four standard metrics, namely *accuracy, precision, recall and F-measure*. Accuracy is simply the ratio of correctly labeled messages to the size of test set, precision is the ratio of the true positives to the sum of true positives and false positives, recall is the ratio of true positives to the sum of true positives and false negatives, and finally, F-measure is the harmonic mean of precision and recall and captures their trade-off.

5.3 Methodology

5.3.1 Research question (RQ) 1

Datasets for investigating RQ1 include *Nepal* and *Kerala* since *Macedonia* does not have a large unlabeled corpus available, which is an assumption made per RQ1. Recall that we used stratified random sampling to split the labeled data for each dataset into training (90%) and test (10%) sets. Of the 90% training set, a further split was done, with 90% kept for 'training' and 10% for setting optimal weights for the 3 linear classifiers[13] trained in Section IV. To account for the effects of randomness, each experiment was conducted across ten trials, with averages reported on all four metrics

[11] Note that the labels are all manually determined and hence, precise; the active learning was only used to *suggest* 'ambiguous' instances from the large unlabeled pool of tweets for manual labeling, not to do the labeling itself (which by definition it cannot, due to the ambiguity inherent in the 'borderline' tweets that we retrieve using the active learning).

[12] Accessed at https://hub.docker.com/r/ppplinday/emergence-detection

[13] The hyperparameters of the linear regression itself were optimized through fivefold cross-validation on this 'inner' (i.e., 90% of the original 90% training set) training set.

Table 4 Results investigating RQ1 on the Nepal and Kerala datasets

System	Accuracy	Precision	Recall	F-measure
(a) Nepal				
Local	63.97%	64.27%	64.50%	63.93%
Manual	64.25%	**70.84%****	48.50%	57.11%
Wiki	67.25%	66.51%	69.50%	67.76%
Local-Manual	65.75%	67.96%	59.50%	62.96%
Wiki-Local	67.40%	65.54%	68.50%	66.80%
Wiki-Manual	67.75%	70.38%	63.00%	65.79%
Our approach	**69.25%*****	68.76%	**70.50%****	**69.44%*****
(b) Kerala				
Local	56.25%	37.17%	55.71%	44.33%
Manual	65.00%	47.82%	**55**.77%	50.63%
Wiki	63.25%	42.07%	46.67%	44.00%
Local-Manual	64.50%	46.90%	51.86%	48.47%
Wiki-Manual	62.25%	43.56%	52.63%	46.93%
Wiki-Manual	**68.75%*****	51.04%	54.29%	**52.20%****
Our approach	68.50%	**51.39%*****	52.76%	51.62%

The bold values in each column represent the best results achieved on that column's metric

described previously for all baselines described below and our approach. Among the different machine learning classifiers in the sklearn package tested, the linear regression was found to work well and used as the classifier of choice where applicable.

We use six baselines to evaluate the approach for RQ1 described in Sect. 4. Note that statistical significance is tested using the one-sided Student's paired t-test by comparing the best system (on each metric) against the *Local* baseline, which is a reasonable choice since in a high-supervision (or even normal-supervision) setting, this baseline has been found to perform quite well. Significance at the 90% level is indicated with a *, at the 95% level with a ** and at the 99% level with a *** (Table 3).

5.3.2 Research question (RQ) 2

The protocol for investigating RQ2 is similar to the one for RQ1. We consider three baselines besides our own approach:

Target-only Local (Target Local): This baseline is essentially the *Wiki-Manual* baseline described in the previous section and trained on the target dataset (i.e., no transfer learning is used, and no source is assumed). This baseline is used to illustrate the benefits of transfer learning, since this baseline sets the minimum benchmark that has to be bested by a transfer learning baseline.

Locally Supervised with Source Embedding (Embedding Transform): Similar to our approach on RQ1, manual features, source embeddings and pre-trained Wikipedia embeddings are used to train three classifiers (but on the labeled target domain) and average their probabilities as the final result. While the local embeddings are trained on the source domain (since unlabeled data are not available for the target domain), all classifier training is always done on the target.

Table 5 Results investigating RQ2 using the Nepal dataset as source and Macedonia dataset as target

System	Accuracy	Precision	Recall	F-measure
Local	58.76%	52.96%	59.19%	54.95%
Transform	58.62%	51.40%	**60.32%***	55.34%
Upsample	59.38%	52.35%	57.58%	54.76%
Our approach	**61.79%***	**55.08%**	59.19%	**56.90%**

The bold values in each column represent the best results achieved on that column's metric

Table 6 Results investigating RQ2 using the Kerala dataset as source and Macedonia dataset as target

System	Accuracy	Precision	Recall	F-measure
Local	58.76%	52.96%	59.19%	54.95%
Transform	62.07%	55.45%	64.52%	59.09%
Upsample	**64.90%*****	**57.98%***	**65.48%*****	**61.30%*****
Our approach	62.90%	56.28%	62.42%	58.91%

The bold value in each column represents the best result achieved on that column's metric

Table 7 Results investigating RQ2 using the Nepal dataset as source and Kerala dataset as target

System	Accuracy	Precision	Recall	F-measure
Local	58.65%	**42.40%**	47.47%	36.88%
Transform	53.74%	32.89%	**57.47%***	41.42%
Upsample	53.88%	31.71%	56.32%	40.32%
Our approach	**58.79%**	35.26%	55.89%	**43.03%***

The bold value in each column represents the best result achieved on that column's metric

Table 8 Results investigating RQ2 using the Kerala dataset as source and Nepal dataset as target

System	Accuracy	Precision	Recall	F-measure
Local	60.26%	**61.80%**	59.94%	59.88%
Transform	**61.18%***	61.04%	63.63%	62.08%
Upsample	60.29%	59.44%	**66.02%***	**62.50%***
Our approach	60.06%	59.54%	63.98%	61.64%

The bold value in each column represents the best result achieved on that column's metric

Locally Supervised with Up-sampling and Source Embedding (Upsample): This baseline is the same as Embedding Transform, except to boost the power of the baseline, we upsample the labeled data (in the target dataset) by 6x. Thus, this baseline tries to mitigate source bias and concept drift by giving more importance to the transfer domain. This baseline is also more appropriate for the case where the target training data are extremely limited.

5.4 Results and discussion

5.4.1 Results: RQ1

Table 4 illustrates the result for RQ1 on the Nepal and Kerala datasets. The results illustrate the viability of urgency detection in low-supervision settings (with our approach yielding 69.44% F-measure on Nepal, at 99% significance compared to the local baseline), with different feature sets contributing differently to the four metrics. While the local embedding model can reduce precision, for example, it can help the system to improve and accuracy and recall. Similarly, manual features reduce recall, but help the system to improve accuracy and precision (sometimes considerably). To truly address the urgency problem, therefore, a multi-pronged ensemble approach is justified, as also argued intuitively in Section IV. We also note that the pre-trained Wikipedia embedding model proved to be an important tool in improving the generalization ability of the model and not requiring any labeled or unlabeled data; in essence, serving as a free resource that could be helped to regularize and stabilize models that would otherwise be uncertain in low-supervision settings.

5.4.2 Results: RQ2

Concerning transfer learning experiments (RQ2), we note that source domain embedding model can improve the performance for target model, and upsampling has a generally positive effect (Tables 5, 6, 7,8). As expected, transfer learning performance (RQ2) is generally lower compared to the low-supervision urgency detection on a *single* dataset[14] (RQ1). Note that at least one of the transfer learning methods always bests the *Local* baseline on all metrics (except precision in Table 7, a result not found to be significant even at the 90% level). Our approach shows a slight improvement over the upsampling baseline on two of the four scenarios (Tables 5, 7) by 2–2.7% on the F-measure metric, which shows the diminishing returns from mixing source and target labeled training data. Further improving performance by high margins will require a radically new approach left for future work.

[14] The best F-measure achieved on Nepal in Table 4 was more than 69%, but when using Kerala as source, only 62.5% F-measure could be achieved (Table 8).

6 Conclusion and future work

This paper presented minimally supervised urgency detection approaches for short texts (such as tweets) in the aftermath of an arbitrary humanitarian crisis such as the 2015 Nepal earthquake. The presented systems covered two scenarios that often emerge in the real world. In the first scenario, a small amount (a few hundred messages) of training data labeled as urgent or non-urgent is available, along with a copious background corpus. In the second scenario, similar data are available for a 'source' domain but not for the target domain (expressing a 'new crisis') for which the urgency detection needs to be deployed. As messages are streaming in for this new domain, investigators label a few samples, but cannot rely on the availability of a background corpus since urgency needs to be tagged in real time before the crisis has fully subsided. To accomplish this challenging goal, our approach relies on a simple but robust transfer learning methodology. Experimental results on three real-world datasets validate our methods.

Some of the obvious avenues for future work are to improve the existing approach incrementally by (for example) adding more manual features and using more sophisticated local embedding model, possibly with more advanced tuning of hyperparameters like the learning rate and vector dimensionality. For improving transfer learning, we are considering using a deep learning model with priors to truly leverage the presence of a source, albeit one covering a domain that is different from the target. Deep learning for transfer learning is still in its infancy in the machine learning community, e.g., a recent survey on deep transfer learning (Tan et al. 2018) shows that most current research *'focuses on supervised learning, how to transfer knowledge in unsupervised or semi-supervised learning by deep neural network may attract more and more attention in the future'*. In looking at the references they cite, the effectiveness of deep transfer learning does not seem to have been demonstrated thus far for difficult and irregular social media datasets However, we believe that this presents an opportunity for further study, especially as new and different crises like COVID-19 continue to threaten our way of life at a global scale.

Acknowledgements The authors gratefully acknowledge the ongoing support and funding of the DARPA LORELEI program and our partner collaborators in providing detailed analysis. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the US Government.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aggarwal CC, Zhai C (2012) Mining text data. Springer, Berlin

Alam A, Bhat MS, Farooq H, Ahmad B, Ahmad S, Sheikh AH (2018) Flood risk assessment of Srinagar city in Jammu and Kashmir, India. Int J Disaster Resil Built Environ 9:114

Alam F, Joty S, Imran M (2018) Domain adaptation with adversarial training and graph embeddings. arXiv preprint arXiv:1805.05151

Alam F, Ofli F, Imran M (2018) Crisismmd: multimodal twitter datasets from natural disasters. In: Twelfth international AAAI conference on web and social media

Anderson KM, Schram A, Alzabarah A, Palen L (2013) Architectural implications of social media analytics in support of crisis informatics research. IEEE Data Eng Bull 36:13–20

Arthur R, Boulton CA, Shotton H, Williams HT (2017) Social sensing of floods in the UK. arXiv preprint arXiv:1711.04695

Avvenuti M, Cresci S, La Polla MN, Marchetti A, Tesconi M (2014) Earthquake emergency management by social sensing. In: Pervasive computing and communications workshops (PERCOM Workshops), 2014 IEEE international conference on, pp 587–592. IEEE

Barrenechea M, Anderson KM, Aydin AA, Hakeem M, Jambi S (2015) Getting the query right: User interface design of analysis platforms for crisis research. In: Cimiano P, Frasincar F, Houben G-J, Schwabe D (eds) Engineering the web in the big data era. Springer, Cham, pp 547–564

Burel G, Saif H, Fernandez M, Alani H (2017) On semantics and deep learning for event detection in crisis situations

Burel G, Alani H (2018) Crisis event extraction service (crees)-automatic detection and classification of crisis-related content on social media

Caragea C, Silvescu A, Tapia AH (2016) Identifying informative messages in disaster events using convolutional neural networks. In: International conference on information systems for crisis response and management, pp 137–147

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12:2493–2537

Crooks A, Croitoru A, Stefanidis A, Radzikowski J (2013) # earthquake: Twitter as a distributed sensor system. Trans GIS 17(1):124–147

Dai AM, Olah C, Le QV (2015) Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998

Faruqui M, Tsvetkov Y, Rastogi P, Dyer C (2016) Problems with evaluation of word embeddings using word similarity tasks. arXiv preprint arXiv:1605.02276

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: Proceedings of the 2017 ACM on web science conference, pp 33–42. ACM

Imran M, Castillo C, Lucas J, Meier P, Vieweg S (2014) AIDR: artificial intelligence for disaster response. In: Proceedings of the 23rd international conference on world wide web, pp. 159–162. ACM

Interdonato R, Guillaume J-L, Doucet A (2019) A lightweight and multilingual framework for crisis information extraction from twitter data. Soc Netw Anal Mining 9(1):65

Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759

Kejriwal M, Szekely P (2017) Information extraction in illicit web domains. In: Proceedings of the 26th international conference on world wide web, pp 997–1006. International world wide web conferences steering committee

Kersten J, Kruspe A, Wiegmann M, Klan F (2019) Robust filtering of crisis-related tweets. In: ISCRAM 2019 conference proceedings-16th international conference on information systems for crisis response and management

Keyvanpour M, Zandian ZK, Heidarypanah M (2020) Omlml: a helpful opinion mining method based on lexicon and machine learning in social networks. Soc Netw Anal Mining 10(1):1–17

Klein B., Laiseca X, Casado-Mansilla D, López-de-Ipiña D, Nespral AP (2012)Detection and extracting of emergency knowledge from twitter streams. In: International conference on ubiquitous computing and ambient intelligence, pp. 462–469. Springer

Kogan M, Palen L, Anderson KM (2015) Think local, retweet global: retweeting by the geographically-vulnerable during hurricane sandy. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. CSCW '15, pp. 981–993. ACM, New York, NY, USA. https://doi.org/10.1145/2675133.2675218

Kruspe A, Kersten J, Klan F (2019) Detecting event-related tweets by example using few-shot models

Kumar S, Barbier G, Abbasi MA, Liu H (2011) Tweettracker: an analysis tool for humanitarian and disaster relief. In: ICWSM

Ladner K, Ramineni R, George K (2019) Activeness of Syrian refugee crisis: an analysis of tweets. Soc Netw Anal Mining 9(1):61

Madichetty S, Sridevi M (2019) Disaster damage assessment from the tweets using the combination of statistical features and informative words. Soc Netw Anal Mining 9(1):42

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119

Moro A, Raganato A, Navigli R (2014) Entity linking meets word sense disambiguation: a unified approach. Trans Assoc Comput Linguistics 2:231–244

Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvisticae Investig 30(1):3–26

Nguyen DT, Joty S, Imran M, Sajjad H, Mitra P (2016) Applications of online deep learning for crisis response using social media information. arXiv preprint arXiv:1610.01030

Nguyen DT, Alam F, Ofli F, Imran M (2017) Automatic image filtering on social networks using deep learning and perceptual hashing during crises. arXiv preprint arXiv:1704.02602

Olteanu A, Castillo C, Diaz F, Vieweg S (2014) CrisisLex: a lexicon for collecting and filtering microblogged communications in crises. In: Proc. int. conf. weblogs and social media (ICWSM), Oxford, UK

Palatucci M, Pomerleau D, Hinton GE, Mitchell TM (2009) Zero-shot learning with semantic output codes. In: Advances in neural information processing systems, pp 1410–1418

Palen L, Anderson KM (2016) Crisis informatics?new data for extraordinary times. Science 353(6296):224–225

Palen L, Soden R, Anderson TJ, Barrenechea M (2015) Success & scale in a data-producing organization: the socio-technical evolution of openstreetmap in response to humanitarian events. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems. CHI '15, pp. 4113–4122. ACM, New York, NY, USA. https://doi.org/10.1145/2702123.2702294

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359

Pang B, Lee L et al (2008) Opinion mining and sentiment analysis. Foundations Trends® Inform Retr 2(1–2):1–135

Pedrood B, Purohit H (2018) Mining help intent on twitter during disasters via transfer learning with sparse coding. In: International

conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation, pp 141–153. Springer

Purohit H, Castillo C, Pandey R (2020) Ranking and grouping social media requests for emergency services using serviceability model. Soc Netw Anal Mining 10(1):1–17

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), pp 119–126. IEEE

Rogstadius J, Vukovic M, Teixeira C, Kostakos V, Karapanos E, Laredo JA (2013) Crisistracker: crowdsourced social media curation for disaster awareness. IBM J Res Dev 57(5):1–4

Romera-Paredes B, Torr P (2015) An embarrassingly simple approach to zero-shot learning. In: International conference on machine learning, pp 2152–2161

Sahlgren M (2005) An introduction to random indexing

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web, pp 851–860. ACM

Settles B (2010) Active learning literature survey. University of Wisconsin, Madison 52(55–66): 11

Simon T, Goldberg A, Adini B (2015) Socializing in emergencies?a review of the use of social media in emergency situations. Int J Inf Manage 35(5):609–619

Soden R, Budhathoki N, Palen L (2014) Resilience-building and the crisis informatics agenda: lessons learned from open cities Kathmandu. In: ISCRAM

Starbird K, Muzny G, Palen L (2012) Learning from the crowd: collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions. In: Proceedings of 9th international conference on information systems for crisis response and management, ISCRAM, pp. 1–10

Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. In: International conference on artificial neural networks, pp 270–279. Springer

Uszkoreit H, Xu F, Li H (2009) Analysis and improvement of minimally supervised machine learning for relation extraction. In: NLDB, pp 8–23. Springer

Verma S, Vieweg S, Corvey WJ, Palen L, Martin JH, Palmer M, Schram A, Anderson KM (2011) Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In: Fifth international AAAI conference on weblogs and social media

Vieweg S, Hughes AL, Starbird K, Palen L (2010) Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1079–1088. ACM

Zheng X, Sun A, Wang S, Han J (2017) Semi-supervised event-related tweet identification with dynamic keyword generation. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 1619–1628. ACM

Zhu X (2005) Semi-supervised learning literature survey