Helping Crisis Responders Find the Informative Needle in the Tweet Haystack

Leon Derczynski*
University of Sheffield / IT University of Copenhagen
leod@itu.dk

Kenny Meesters
TU Delft
k.j.m.g.meesters@tudelft.nl

Kalina Bontcheva
University of Sheffield
k.bontcheva@sheffield.ac.uk

Diana Maynard
University of Sheffield
d.maynard@sheffield.ac.uk

ABSTRACT

Crisis responders are increasingly using social media, data and other digital sources of information to build a situational understanding of a crisis situation in order to design an effective response. However with the increased availability of such data, the challenge of identifying relevant information from it also increases. This paper presents a successful automatic approach to handling this problem. Messages are filtered for informativeness based on a definition of the concept drawn from prior research and crisis response experts. Informative messages are tagged for actionable data – for example, people in need, threats to rescue efforts, changes in environment, and so on. In all, eight categories of actionability are identified. The two components – informativeness and actionability classification – are packaged together as an openly-available tool called Emina (Emergent Informativeness and Actionability).

Keywords

informativeness, twitter, social media, actionability, information filtering

INTRODUCTION

A key aspect of dealing with the aftermath of disasters and critical events is to get a quick situational understanding of the situation. This situational understanding enables responders to not only understand the gravity of the situation and the needs of the affected community, but also the operational circumstances and the availability of resources. However, during a crisis situation getting a coherent understanding of the situation can prove challenging, due to its chaotic nature. Therefore crisis responders are continuously looking for key pieces of information to build, implement and improve their response (Harrald 2006).

In recent years, the advent of information technology has provided crisis responders with new tools and systems to collect information directly from the communities affected by crisis. At the same time, with the rise of social media and the profusion of mobile technology the same communities are also increasingly empowered – and at times incentivized – to share their data through an increased number of channels. These developments provide crisis responders with access to more data than ever before, directly from the communities affected by the crisis (S. Vieweg et al. 2010).

However, this data surge has given rise to new challenges. Where before the key challenges were to gather sufficient data to build a situational understanding, the main challenges now are to identify the relevant messages (Comes, Meesters, et al. 2017). While many responding agencies have confirmed the usefulness of social media and other

*corresponding author

forms of electronic information gathering, it remains challenging to integrate an effective workflow in short-burning crisis situations (Meesters et al. 2016). Currently, organizations often rely on external volunteers, such as the Standby Task Force in the wake of the 2010 Haiti Earthquake (Patrick et al. 2011; Rencoret et al. 2010). However, not all organizations have the capacities or knowledge to process the large amounts of data available to them in the search for relevant messages.

Information filtering is critical to crisis workers dealing with streams of information related to emerging and ongoing situations (S. E. Vieweg 2012; Temnikova et al. 2015). Today, messages in many forms arrive constantly in volumes too large for humans to effectively prioritize in real-time. In this research, we aim to provide automated methods for filtering this information so that crisis workers can effectively find potentially important messages and perform information triage.

Crisis worker flow typically begins with defining selectors that determine what information is retrieved; these could be keywords, particular sources, GPS coordinates, and so on. For example, to get content about hurricane Irma, one could capture just messages that contain “irma" and originate within the monitored area. This is the first pass at reducing the volume of information, but even with this, the volume of information balloons rapidly as events progress; data for Typhoon Yolanda in 2013 goes from 182 tweets in the first three days to 1778 on the fifth day and 4385 a week later; the west Texas explosion generated 10K tweets on the first day, going from 128 tweets in hour 1 to 2456 in hour 4.[1] And indeed these selectors do not guarantee that the incoming data is good; many messages can be irrelevant (Homberg et al. 2014).

These challenges are however not only related to natural disasters, but also extend to other disruptions and incidents faced by communities around the world. For example, in a recent iHub project exercise monitoring elections in Kenya, workers found roughly 90% of the messages captured in the message monitoring Ushahidi platform were irrelevant[2] – not only due to their content, but also due to missing content, repetitive messages, and so on. In this case the messages were deliberately solicited: people were encouraged to post their own observations (via various channels). Despite this ’top-down’ approach, the collected data still resulted in the high percentage of irrelevant messages. In sudden onset disasters, the number of messages and noise is potentially much higher. Methods to ease the impact of high data volume are therefore still pressing.

These developments and trends show that a tool to reduce the amount of data seen by humans and filter to just the relevant messages is not only necessary but would improve the usefulness of the data available to responders. The development of auto-assessment of informativeness is the first challenge that needs to be addressed towards a system of automated identification of relevant messages for crisis responders.

In this paper, we describe the openly-available Emina (Emergent Informativeness and Actionability) tool, which we have developed to address these issues as part of the EU COMRADES project.[3]

RELATED WORK

The problem of assessing informativeness automatically, or of extracting salient posts from intense social media streams, is not unfamiliar. Prior work has addressed informativeness extraction and coarse classification using DNNs (Nguyen et al. 2016). Over this, we add many categories of actionability, while also refining the informativeness decision using a new dataset and training target. Other efforts have addressed the data scarcity problem by recycling prior data (Imran, Mitra, and Srivastava 2016) over a two types of event (earthquakes and floods), or focused on extracting individual self-contained documents that are highly relevant, as exemplars of the ongoing crises or high-quality messages that should be responded to (Imran, Elbassuoni, et al. 2013).

DEVELOPMENT APPROACH

Our approach began by developing a definition for informativeness. With this, we can gather and annotate data for its informativeness, using a crowdsourcing task, where readily-available workers are drawn from a pool over the internet to quickly process the kinds of tasks that require human intelligence. This produces our training set. This training set is used to train a machine learning system, which learns mathematical rules about the text in the training set that help it predict the informativeness of future, previously-unseen messages, mimicking a deployed situation. The same general approach is taken to learning actionability.

[1] From CrisisLex data (Olteanu, Castillo, et al. 2014)

[2] See https://www.ushahidi.com/blog/2017/08/10/ushahidis-global-uchaguzi-partnership-crowdsourcing-kenyan-election-incidents-shows-a-mostly-orderly-election

[3] See http://comrades-project.eu

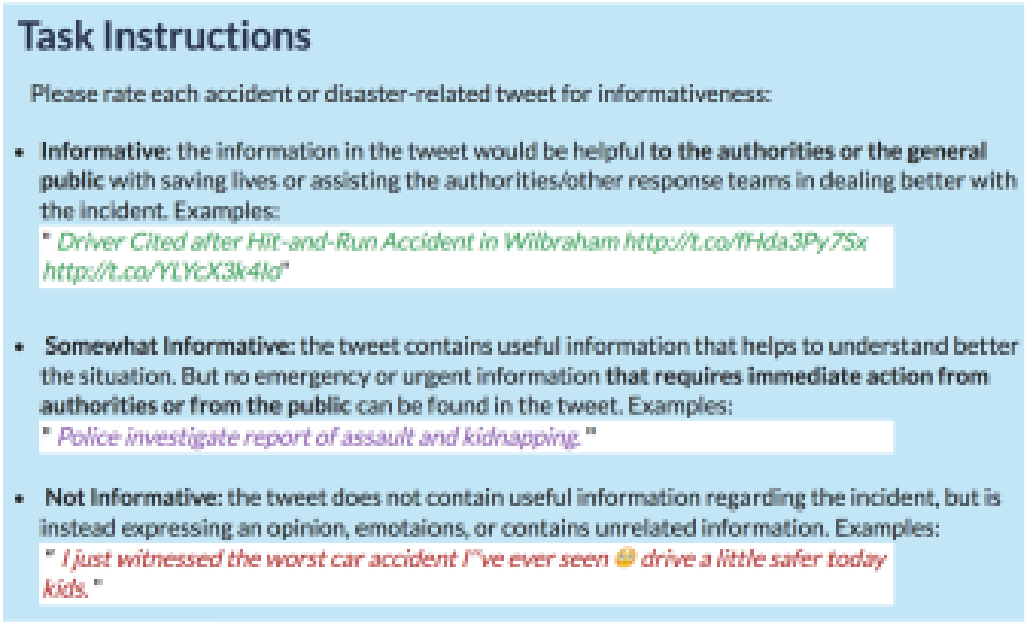

Figure 1. How the informativeness task is presented to crowd workers

The developed method for assessing the informativeness and filtering the messages is divided into two stages. Firstly, incoming messages are labelled for informativeness by a machine learning system. Messages deemed to be uninformative are discarded. The sensitivity of this system can be tuned. For example, in situations where it is critical to process every potentially important message, this can be achieved, at the cost of some irrelevant messages coming in. In other situations, where it's most important to keep up to date, and only summary information is required, then all irrelevant messages can be discarded at the risk of some potentially relevant ones being missed.

The second stage is to determine what kind of action can be taken based on the message. This is termed *actionability classification*. This paper describes a novel schema for actionability classification. The schema is prototyped, updated, and then used to construct a dataset. A machine learning system is then trained to classify incoming messages according to their actionability, based on this dataset.

PREDICTING INFORMATIVENESS

Discarding uninformative data is critical when one aims to process crisis-relevant messages in real time. Typically, when capturing messages related to a crisis, any potential selector (e.g. geographical area, hashtag) yields a high volume of available content, via e.g. Tweets or SMS. However, only a limited part of this will in fact be relevant to the crisis and also informative.

Therefore, the first step in dealing with any information stream related to a crisis is to differentiate between informative and non-informative content. The two major technical challenges here are first in developing a computationally operationalised definition of informativeness in crisis situations, and secondly in creating a technical solution that can discriminate based on informativeness in the extraordinarily broad range of not only crisis situations but also the social and other message media used around them.

Defining Informativeness

A key challenge in developing an automated approach for identifying relevant messages is to describe 'relevant' in a manner that enables it to be transformed into an automated process. However, identifying messages which are relevant, or informative (having an information value), is often both heavily context-dependent and also subjective. Furthermore, informativeness patterns (and associated relevancy of information) may differ from one type of crisis to another (Olteanu, S. Vieweg, et al. 2015). Because we rely partially on human-annotated data to train our system, the definition of informativeness must additionally both provide useful distinctions for the end users of the tool (the crisis workers) and be as easy as possible for human annotators to understand and apply to the data they are given. Previous work has shown that the quality of crowdsourced data is heavily dependent on both the simplicity and clarity of the task and definition (Sabou et al. 2014), and this data is critical for the development of the tool.

These issues makes it challenging both to capture good examples of informativeness, and to develop a definition for it that can be used for crisis analysis. For this scenario, the notion of informativeness ideally needs to reflect the extent to which information captured during a crisis is useful to responders; or at least whether or not a message is related to the crisis. Rather than build a definition from first principles, we instead took a top-down approach, considering what kinds of messages in practice would be useful, and generalising from that. To arrive at a definition, we drew on a range of existing sources and efforts. These included third-party resources from established authorities in the field of information management for crisis response, such as CrisisLex (Olteanu, Castillo, et al. 2014), the OCHA Guiding Principles (Van de Walle et al. 2008), and the World Disaster report, as well as academic sources such as the DERMIS principles (Turoff et al. 2004), and field research in Nepal (Baharmand et al. 2016). Finally, we included research carried out within the COMRADES project, including monitoring exercises from iHub and input from Delft University of Technology (Comes, Toerjesen, et al. 2016).

From these resources, and after some trial and error testing different definitions on a number of volunteer annotators and getting feedback, we created a three-level definition of informativeness. It became clear that informativeness often hinges on how urgent information is, so this concept is reflected in the levels:

- **Informative:** the information in the message would be helpful with saving lives or assisting the authorities/other response teams in dealing better with the incident.
- **Somewhat Informative:** the message contains useful information that helps to understand better the situation. But no emergency or urgent information can be found in the tweet.
- **Not Informative:** the message does not contain useful information regarding the incident, but is instead expressing an opinion or contains unrelated information

Because we crowdsource the work to human annotators, we were mindful of how task design and framing could affect results: taking care of cognitive biases is paramount to achieving good results (Sabou et al. 2014; Derczynski, Bontcheva, et al. 2016; Eickhoff 2018). Specifically, we wanted annotators to think carefully about mid-cases. Our initial tests showed that a binary decision (informative or not) was too rigid; on the other hand, giving annotators too many categories made it hard to find the boundaries between them. A sample task is shown in Figure 1.

Data Collection

The text data used in this stage of the project was part of the dataset released in (Schulz et al. 2017). This dataset includes over 10K tweets that were selected and labelled for the domain of incident detection of 10 different cities. The tweets were already labelled by crowdsourced annotators, and classified. The first dataset has two classes (related and not related to the incident); the second has four classes (crash, fire, shooting, and not related). Note that there is a distinction between related and informative: related messages are not necessarily informative, but unrelated messages are definitely not informative. We thus retrieved 600 tweets from the first dataset taking only those labelled as related. Following this, we filtered our dataset to remove all retweets, as well as tweets with redundant content (i.e. subjectively carrying the same information or describing the same fine-grained event).

Annotators on CrowdFlower[4] were then instructed to label these documents according to the three informativeness categories listed above. Five crowd worker judgments were collected for each tweet. Crowd responses were then adjudicated taking the most popular category. In general, total agreement was higher than 70% in all three groups, and over 80% for the “not informative" category. The resulting annotated data has been made publicly available as the COMRADES Crowdsourced Informativeness Dataset (CCSID) (Qarout et al. 2018).

Supplemental Data

At least a few thousand examples are required to train most machine learning systems. For extra data, we used CrisisLex (Olteanu, Castillo, et al. 2014), which contains around 60,000 tweets. These are labelled for informativeness, though not as stringently as CCSID; messages are coded as either “Informative", corresponding to the union of CCSID’s first two options, or “Non-informative", corresponding to the third. In any event, it comprises 32K positive and 28K negative examples. While this dataset is not as reliable (for example, there are anecdotal cases of useless “informative" tweets in it), the effect of this is diminished by balancing data and exploiting loss functions during training. This helps us to get high quality from the CCSID while leveraging the bulk of the CrisisLex annotations, as described below.

[4]http://www.crowdflower.com/

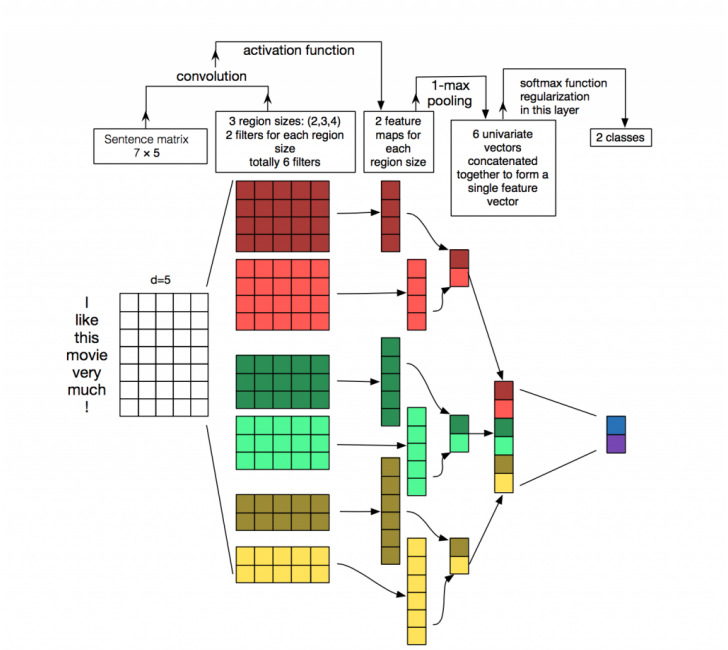

Figure 2. The convolutional neural network architecture used (from Zhang et al. 2015)

Machine Learning Approach

Having the data, the next step is to learn what informative messages generally look like. A machine learning algorithm will pick up patterns in the data we present it with, and try to generalise from that what signals an informative message. Then, when applied to new messages, it should be able to make some decision about what is and is not informative. In this instance, because we are dealing with highly diverse text and have a reasonable amount of data, we choose a convolutional neural network, which ought to be able to infer directly from text without additional feature extraction.

Because we need to leverage the data from CrisisLex as well as CCSID, we can label a maximum of two classes, as CrisisLex does not have the middle “informative but not urgent" label. Therefore, for this system, we reduce the CCSID categories from three to two by combining “informative and urgent" with “informative but not urgent".

Convolutional networks work by repeatedly applying a “window" to a numerical representation, and combining all the values under that window. For text, those numbers come from a word embedding. We followed the convolutional text classification network of Zhang et al. 2015 to determine informativeness (Figure 2).

It is important to know how good a machine learning model is, so that we know when to stop training. Sometimes, training for more time does not give a better result. This is because some systems match the training examples too closely, and so will not operate well in a new environment. For example, if our training set only took examples from shootings, and the term “AK47" occurred often in this data, then a system might end up thinking that highly specific mentions of “AK47" indicate informativeness, instead of phrases like “wounded". This is called overfitting.

The quality of a machine learning model is judged in two ways. First, we can measure how well the trained model works on the data that is used to train it (the “training set"). This is like measuring how well the system has learned the data to which it has been exposed. The metric used for this is “training loss", and lower is better. The second way is to measure the model’s performance on some other, held-out data (a “validation set"). A good score on this indicates that the model is learning to predict things beyond the data it has already seen, i.e. the model is generalising. This is called “validation loss", and again, a lower value is better.

Typically, one takes a dataset, and splits off a part which becomes the validation set. During training, training loss decreases, as does validation loss. When training loss becomes lower than the validation loss, the model has started to predict the training data better than other data, meaning it has started overfitting. So, the model selected is the one just before this crossover of validation and training loss, with a minimum validation loss.

In the instance of our informativeness classifier, we have a somewhat limited dataset (large scale machine learning text datasets typically comprise millions of examples). In addition, some of the training data (ie. from CrisisLex) is not a precise match for the kinds of scenarios we envisage COMRADES outputs being deployed in. This means that

measuring loss against CrisisLex alone might gives a system that is very good for the kind of scenarios contained in this dataset, but does not generalise well across crises.

To remedy this, we use part of the CCSID in the validation set, and give it a much larger proportion in the validation set than the test set. Specifically, the validation set is comprised of 300 CCSID instances and 300 drawn from CrisisLex (150 each of informative and non-informative). The remainder of the data, some 60,000 examples, forms the training data. This means that scoring well on the more-varied CCSID, which matched COMRADES goals well, is very important, despite it only comprising a smaller proportion of the training data. In this way, we can select a model for informativeness classification that uses all the data we have available and is optimally likely to score well in real-world deployment.

Evaluation

This training and validation partition were used with the configuration specified above. The final scores in this case are an accuracy of 92% and an F1 of 91%. This is over the validation dataset, which represented both the definition of informativeness we wanted to capture and also the diversity present in social media. The F1 indicates that we achieve an excellent balance in errors between false positives (i.e. uninformative messages slipping through) and false negatives (informative messages being missed). We attribute the high performance of this system to the careful annotation process, as well as the volume and range of documents available for training. Additionally, we purposefully created a system aimed at high performance on this data type.

ACTIONABILITY

Having filtered out uninformative content, the next stage is to assist crisis workers in information processing by categorizing the remaining messages. This could be used to select individual messages and hand them over to a specialist team, as well as to get a good overview of the crisis situation or detect emerging events.

Some former work has addressed this. Recently, the CREES system (Burel et al. 2017) used the CrisisLex categories and large amount of data there to do some filtering of information into different categories. However, not all these were actionable, and while there is some overlap, the categories are either not directly useful when determining actionability, or do not provide a sufficient amount of granularity. Nevertheless, with its large amount of data, CREES achieves good performance. Earlier work on determining actionability in crises was effective though studied over a single crisis (Munro 2011). We go further and train on multiple crises, allowing a more general impression of how actionability can look over a broad range of contexts.

Defining Actionability

Our requirements were that a relatively small number of categories should be selected, and all of these at a single, top level. A hierarchical system is complex and hard to build; the added complexity here incurs cost and quality risk, and for a real-time prototype system, it is better to provide something that works. Further subcategorization could always be efficiently provided later if necessary.

Additionally, the categories should correspond to issues relevant to the type of crisis tracked by COMRADES and envisaged for use by project stakeholders. This means some categories in other datasets feature less in our work. For example, media opinion of the crisis response is less important in our use cases than trapped and injured people. The source categories are given here; firstly, for the MIT Humanitarian Response Lab's schema, as in Figure 3.[5]

[5]See https://humanitarian.mit.edu/ (MIT HR is also the source of the schema diagram)

Figure 3. A schema from the MIT Humanitarian Response Lab for prioritising hurricane response

Secondly, from CrisisNLP (Imran, Mitra, and Castillo 2016):

1. Injured or dead people: Reports of casualties and/or injured people due to the crisis
2. Missing, trapped, or found people: Reports and/or questions about missing or found people
3. Displaced people and evacuations: People who have relocated due to the crisis, even for a short time (includes evacuations)
4. Infrastructure and utilities damage: Reports of damaged buildings, roads, bridges, or utilities/services interrupted or restored
5. Donation needs or offers or volunteering services: Reports of urgent needs or donations of shelter and/or supplies such as food, water, clothing, money, medical supplies or blood; and volunteering services
6. Caution and advice: Reports of warnings issued or lifted, guidance and tips
7. Sympathy and emotional support: Prayers, thoughts, and emotional support
8. Other useful information: Other useful information that helps understand the situation
9. Not related or irrelevant: Unrelated to the situation or irrelevant

We settled on a blended prototype set of top-level actionable information types. Unlike CrisisNLP, this set imposes no specific ordering or prioritization on the data. This ordering is not applicable in all crises, nor indeed over time in the same crisis. For example, dead bodies are not a priority compared with the trapped and injured, on day zero; but on day 3-4, they may well be.

To avoid colouring user judgments, we took a conscious decision to make all tools as agnostic as possible, cutting out any biases that we noticed. This has the goal of placing all decisions about response and action on the analyst, instead of providing nudges about what has higher or lower rankings (as one implicitly does by ordering actionability types).

- A specific resource, where the kind of need is given explicitly.
- Mentions of some group that's responding or aiding (e.g. volunteers, military, government, businesses, aid agencies).
- Threats to the general crisis response. Weather warnings, fires, military action etc.
- Changes in accessibility - limits to access, or other changes in how much transport can get through.
- Damage to infrastructure, livelihoods etc.
- Mention by name of the geographic areas affected
- Changes in the local environment (weather, hazards, etc.); e.g., a storm is intensifying
- General reporting about the rescue effort (from the media or the public)
- Opinion or individual message.

Note that some categories, such as those around Needs, are collapsed into one. Following this example, we first should identify content about needs, and analysis of the exact type of information – for example, the resource required, or the location – can be postponed (Munro 2011).

Category	Count
Needs	22
Response groups	99
Threats to response	195
Changes in accessibility	19
Damage to livelihoods	62
Mention of affected areas	400
Changes in environment	40
General reporting	83
Personal opinion	423

Table 1. Actionability data

Gathering Actionability Data

Labelling qualitative data with quantitative, hard categories is a difficult process, with some margin of error. The standard of this process directly reflects the quality of the resulting dataset and the performance of machine learning systems trained on them.

Humans typically perform rather poorly in data labelling when faced with more than seven labels, and best when faced with two (Sabou et al. 2014). As we have nine categories, the set must be decomposed somehow. As with the informativeness task, we trialled various different combinations and setups using in-house volunteer annotators. We had best results from splitting types into two groups, and providing the “personal opinion" option alongside both.

To maintain quality in crowdsourcing, a first portion of the data was annotated by experts who had experience in the goals of the data, crisis response, and linguistic annotation. After settling discrepancies, this pilot part of the data was converted to 118 “gold" messages which would be used as test examples for crowd workers; for each gold example, the correct answers are pre-filled, and then used to give feedback on the job as crowd workers proceed. The test data gives crowd workers a safe environment in which to learn how to annotate, while seeing real examples of correctly-labelled data. This follows best practices in crowdsourced annotation of corpora (Sabou et al. 2014).

We combined tweets from all days of Hurricane Irma[6] with data from CrisisLex, taking from the latter only crises of natural origin, in order to filter out short-term anthropogenic events like shootings. These were then filtered for informativeness based on the system introduced above, so that only informative tweets were scanned for actionable information. In total, 1350 tweets were labelled according to the actionability criteria indicated above, distributed as shown in Table 1.

Note that an individual tweet can contain more than one kind of information (or no relevant information at all). Therefore, the totals above do not sum to the dataset size. This comprises our dataset of actionable messages, annotated by a mixture of experts and crowd workers that have been trained with expert data.

Extracting Actionable Content

Having now a set of examples of actionable and non-actionable data, it is possible to train a system to triage messages according to actionability.

Deep learning requires large amounts of data; its advantage lies in a better ability to leverage better data. In this situation, however, we are faced with a paucity of examples from each actionability class, which is tough for deep learning. Indeed, applying a convolutional text kernel as with earlier informativeness analyses led to very poor results, with a F-score under 0.1 in almost every case. We attribute this to the extreme data sparsity.

Further, data where there is a strong imbalance is harder for most machine learning algorithms to learn to predict. In this case, our dataset of over 1000 instances often contains only a minority of positive examples, complemented by an overwhelming majority of negative examples. This is a difficult learning environment: for effective learning the number of positive and negative examples should be roughly equal.

We cast actionability extraction as a multi-class labeling problem. That is, a message can have more than one kind of actionable information at a time. For example,

There are people camped out between Lincoln and fifth who need water

[6]Kindly provided by PushShift – http://pushshift.io

Class	Accuracy	F1	Recall	Baseline F1
Needs	66.83%	57.50%	63.01%	45.21%
Response groups	63.13%	62.94%	62.63%	40.18%
Threats to response	57.18%	57.72%	58.46%	33.70%
Change in accessiblity	57.89%	57.89%	57.89%	18.18%
Damage to infrastructure, livelihoods	56.45%	54.24%	51.61%	20.16%
Geographic names	56.63%	57.84%	59.50%	45.23%
Changes in environment	50.00%	47.37%	45.00%	23.58%
Reporting about the rescue	57.83%	58.82%	60.24%	13.95%
Personal opinions	66.59%	63.43%	57.94%	42.27%

Table 2. Performance of actionability detectors

contains both a need (type A) and a geographic location mentioned by name (type J) – so it's labeled [A,J]; hence, multi-class. We construct a solution based on many binary classifiers, one for each actionability type. Each classifier will attempt to predict whether or not the message contains a single type of actionability data, and a single classifier is dedicated to each kind of actionable information. Finally, the results are pooled to form an aggregate labeling of the message.

For feature extraction, following Aker et al. 2017, we extracted from the data a set of keywords for each actionable information type, and also induced word embeddings over recent social media data as a continuous representation of individual word types. These were taken from 20 million tweets from 2016, filtered for English with langid.py (Lui and Baldwin 2012) randomly selected for variation in time (across seasons and across time of day), processed with GloVe (Pennington et al. 2014), and having 25 dimensions (a low dimensionality is less prone to errors with this relatively sparse training data). Documents are converted to data by splitting them into words using a social media tokenizer (O'Connor et al. 2010) and then, for each word in the keyword list, measuring the proportion of words in the document that are similar to the keyword. Similarity is defined by cosine vector proximity, with a cut-off of 0.45. That is to say, document words whose vectors have a cosine with the keyword under 0.45 are disregarded. Additionally, each document word is scaled by its score; so, if a ten-word document has one similar keyword, and the similarity is 0.6, the overall value for this keyword is 0.06 (10% × 0.6). This feature extraction converts documents into a one-dimensional vector with a value for each keyword.

Sample keywords for two of the actionability types are:[7]

Personal opinion: `[i people us all me please good will trump your pray praying toll concerned omg worried god]`

Changes in accessibility: `[G accessibility street bridge blocked derailment collapse close flooded closed careful cancel cancelled canceled avoid alert affect advised]`

These terms are those likely to be indicative of a particular class. They are extracted by finding the most discriminative words for each class through a simple Bayesian analysis. The full set of keywords and actionability types are downloadable from the Emina distribution and can be edited to suit new situations.

We needed a classifier that was rapid, could learn with with a low amount of data, and that did not select a liner decision boundary. To this end, the RBF SVM fit our needs. In addition, we reduced the "negative" data (i.e. the non-actionable cases) down to match the number of positive examples for each class, randomly removing examples.

Evaluation

We measured both the accuracy of the system (i.e. the proportion of examples it got right) as well as its F1 score and its recall (i.e. the proportion of known actionable messages that were retrieved by the classifier). Recall was particularly important here as precise systems (i.e. those that return only actionable information) were not considered as useful as those with good recall; for us to correctly pick up two instances of people in need and ignore four hundred would still have high precision – when no false need messages were returned (i.e. those where there is no need expressed) – but poor impact on crisis analysis and response.

[7] The full set is provided with the Emina toolkit

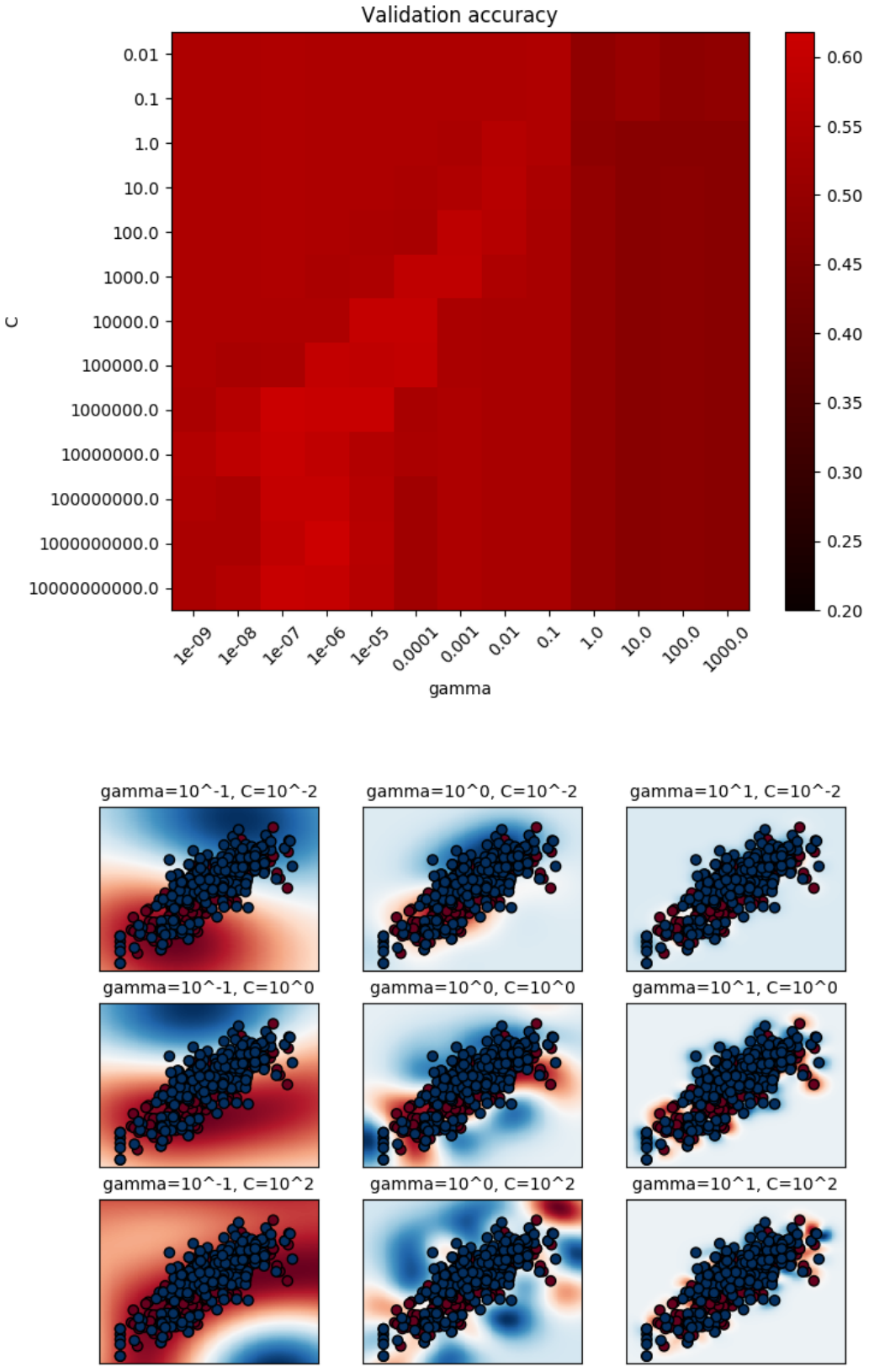

Figure 4. Tuning parameters for the RBF SVM

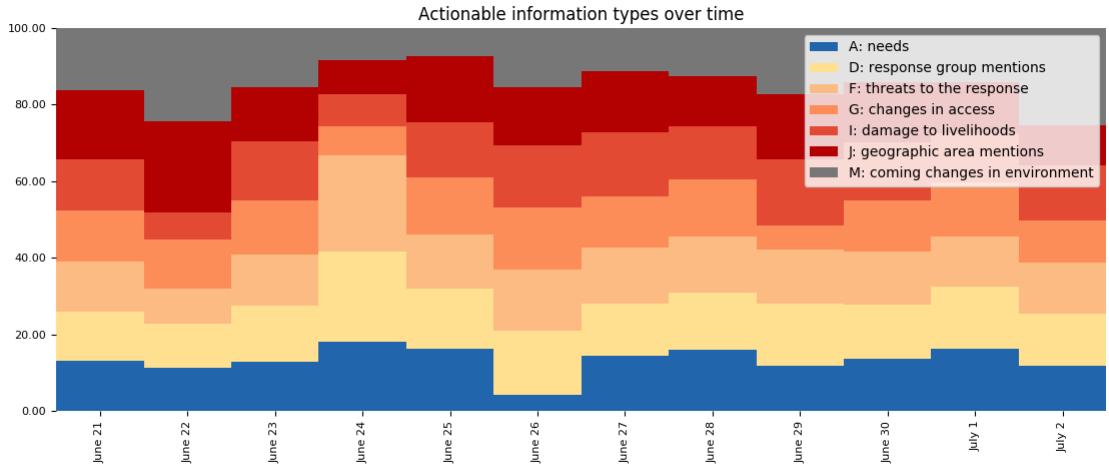

Figure 5. Types of actionable data during the 2013 Alberta floods

Radial Basis Function (RBF) SVM had excellent performance in some categories and good performance in others. Other classifiers did best in some areas but much worse in others. RBF SVM's overall performance indicates it is stable for this task especially given that we have relatively small datasets and the way messages are written varies very wildly across crises, even more so than to be expected in social media (Derczynski, Maynard, et al. 2013).

The RBF SVM works by exploring the space out from each example, to see where boundaries lie. This effectively tells us the impact or sensitivity of each keyword, in the context of other keywords. It seems that such an approach is more effective than finding linear separators in such a data-imbalanced, high-variance environment.

RBF is tuned with two parameters, C and gamma. The gamma parameter defines the influence that a single data point has. C trades off error against simplicity – a very complex system might make few errors, but may be fragile, having very poor performance on new data. The behavior of an RBF SVM can be highly sensitive to gamma; large values give individual examples power to distort the system's behavior. To best tune our system, we searched for parameters (Figure 4. Results are for category G, access to infrastructure, though simplified for visualization to the first two keyword parameters.

The plots show that performance is best with a high C and a low gamma – that is, we should not fit the data too tightly to either of these points. In the end, C=20 and gamma=3 worked for the general case. Overall results for RBF SVM are given in Table 2. These are compared with a keyword-match baseline, which matches texts if any of the keywords for a given actionability type are present.

This actionability extraction, coupled with the informativeness filtering, is released as an open source tool Emina at https://github.com/GateNLP/emina.

PROFILE OF A CRISIS

Using both informativeness filtering and actionability classification, we are able to see how the kind of actionable information in a crisis situation develops over time. To this end, we applied informativeness filtering and then actionability extraction to messages around individual crises in prior datasets.

Visualization

To measure the balance of actionable information over time, we bucketed messages according to their timestamps and then calculated the proportion of each actionability type. The results are shown in a proportional stacked bar chart. That is to say, for each time slot, the chart shows the proportion of actionable information during that period. If the dominant information type is "needs", then this second will have a larger proportion than the others, regardless of volume. An example for the 2013 Alberta floods is shown in Figure 5. This kind of visualization is an accessible summary of the profile of the crisis over time.

CONCLUSION

Filtering out uninformative messages is crucial to effective use of human monitors of electronic messages during a crisis. The majority of inputs in recent crises are irrelevant, and systems that preprocess and reduce the volume

of irrelevant messages become invaluable. We developed a pair of techniques that first remove the bulk of irrelevant messages, and then attempt to select particular kinds of actionable messages, according to some top-level actionability categories.

We were able to extract informative messages at over 90% accuracy, incorporating both the accurate selection of informative content and also avoiding extraction of non-informative content. Furthermore, we were able to recall between 50% and 70% of actionable information over eight individual information classes. The results presented in this paper are a first step to examine the feasibility of automatically identifying informative, and actionable messages such as those present from social media sources during a crisis event. While the results in this paper are promising, further research is needed to improve this approach. More importantly, given the importance of data and responsibilities of crisis responders, additional research is needed into the risks and ethics associated with automatic data processing. With automated services and tools, like Emina presented in this paper, we gain the potential to support crisis responders, digital volunteers and communities themselves in processing large volumes of data.

ACKNOWLEDGMENTS

This work was supported by funding from the European Commission through Horizon 2020 grant agreement 687847, COMRADES, and grant agreement 654024, SOBIGDATA. The authors are grateful to CrowdFlower and Rob Munro for their continued support and the resources generously donated which enabled creation of these datasets, as well as to Pushshift.io and Jason Baumgartner, who helped provide the Irma data rapidly during and after the crisis. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

REFERENCES

Aker, A., Derczynski, L., and Bontcheva, K. (2017). "Simple Open Stance Classification for Rumour Analysis". In:

Baharmand, H., Boersma, K., Meesters, K., Mulder, F., and Wolbers, J. (2016). "A multidisciplinary perspective on supporting community disaster resilience in Nepal". In: *Proc. ISCRAM*.

Burel, G., Saif, H., and Alani, H. (2017). "Semantic Wide and Deep Learning for Detecting Crisis-Information Categories on Social Media". In: *International Semantic Web Conference*. Springer, pp. 138–155.

Comes, T., Meesters, K., and Torjesen, S. (2017). "Making sense of crises: the implications of information asymmetries for resilience and social justice in disaster-ridden communities". In: *Sustainable and Resilient Infrastructure*, pp. 1–13.

Comes, T., Toerjesen, S., and Meesters, K. (2016). *D2.1: Requirements for boosting community resilience in crisis situation*.

Derczynski, L., Bontcheva, K., and Roberts, I. (2016). "Broad Twitter Corpus: A Diverse Named Entity Recognition Resource". In: *Proc. COLING*, pp. 1169–1179.

Derczynski, L., Maynard, D., Aswani, N., and Bontcheva, K. (2013). "Microblog-genre noise and impact on semantic annotation accuracy". In: *Proc. Conference on Hypertext and Social Media*. ACM, pp. 21–30.

Eickhoff, C. (2018). "Cognitive Biases in Crowdsourcing". In: *Proc. WSDM*.

Harrald, J. R. (2006). "Agility and discipline: Critical success factors for disaster response". In: *The annals of the American Academy of political and Social Science* 604.1, pp. 256–272.

Homberg, M. van den, Meesters, K., and Van de Walle, B. (2014). "Coordination and Information Management in the Haiyan Response: observations from the field". In: *Procedia Engineering* 78, pp. 49–51.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). "Extracting information nuggets from disaster-related messages in social media". In: *Proc. ISCRAM*.

Imran, M., Mitra, P., and Castillo, C. (2016). "Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages". In: *Proc. LREC*.

Imran, M., Mitra, P., and Srivastava, J. (2016). "Enabling Rapid Classification of Social Media Communications During Crises". In: *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 8.3, pp. 1–17.

Lui, M. and Baldwin, T. (2012). "langid.py: An off-the-shelf language identification tool". In: *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics, pp. 25–30.

Meesters, K., Beek, L. van, and Van de Walle, B. (2016). "#Help. The Reality of Social Media Use in Crisis Response: Lessons from a Realistic Crisis Exercise". In: *Proc. HICSS*. IEEE, pp. 116–125.

Munro, R. (2011). "Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol". In: *Proc. CoNLL*. ACL, pp. 68–77.

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

O'Connor, B., Krieger, M., and Ahn, D. (2010). "TweetMotif: Exploratory Search and Topic Summarization for Twitter". In: *Proc. ICWSM*, pp. 384–385.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises". In: *Proc. ICWSM*.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to expect when the unexpected happens: Social media communications across crises". In: *Proc. CSCW*. ACM, pp. 994–1009.

Patrick, J. et al. (2011). "Evaluation insights Haiti earthquake response emerging evaluation lessons". In: *Evaluation Insights*.

Pennington, J., Socher, R., and Manning, C. (2014). "GloVe: Global vectors for word representation". In: *Proc. EMNLP*, pp. 1532–1543.

Qarout, R., Maynard, D., Derczynski, L., and Bontcheva, K. (2018). *COMRADES Crowdsourced Informativeness Dataset (CCSID)*. https://figshare.com/articles/COMRADES_Crowdsourced_Informativeness_Datased_CCSID_/5787693/2.

Rencoret, N., Stoddard, A., Haver, K., Taylor, G., and Harvey, P. (2010). "Haiti Earthquake response context analysis". In:

Sabou, M., Bontcheva, K., Derczynski, L., and Scharl, A. (2014). "Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines." In: *Proc. LREC*, pp. 859–866.

Schulz, A., Guckelsberger, C., and Janssen, F. (2017). "Semantic Abstraction for generalization of tweet classification: An evaluation of incident-related tweets". In: *Semantic Web* 8.3, pp. 353–372.

Temnikova, I., Vieweg, S., and Castillo, C. (2015). "The case for readability of crisis communications in social media". In: *Proc. WWW*. ACM, pp. 1245–1250.

Turoff, M., Chumer, M., Van de Walle, B., and Yao, X. (2004). "The design of a dynamic emergency response management information system (DERMIS)". In: *JITTA: Journal of Information Technology Theory and Application* 5.4, p. 1.

Van de Walle, B., Van Den Eede, G., and Muhren, W. (2008). "Humanitarian information management and systems". In: *International Workshop on Mobile Information Technology for Emergency Response*. Springer, pp. 12–21.

Vieweg, S. E. (2012). "Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications". PhD thesis. University of Colorado at Boulder.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). "Microblogging during two natural hazards events: what twitter may contribute to situational awareness". In: *Proc. CHI*. ACM, pp. 1079–1088.

Zhang, X., Zhao, J., and LeCun, Y. (2015). "Character-level convolutional networks for text classification". In: *Proc. NIPS*, pp. 649–657.