Verifying Baselines for Crisis Event Information Classification on Twitter

Justin Michael Crow
TAG-lab*
University of Sussex†
jmcrow@protonmail.com

ABSTRACT

Social media are rich information sources during crisis events such as earthquakes and terrorist attacks. Despite myriad challenges, with the right tools, significant insight can be gained to assist emergency responders and related applications. However, most extant approaches are incomparable, using bespoke definitions, models, datasets and even evaluation metrics. Furthermore, it's rare that code, trained models, or exhaustive parametrisation details are openly available. Thus, even confirming self-reported performance is problematic; *authoritatively* determining state of the art (SOTA) is essentially impossible. Consequently, to begin addressing such endemic ambiguity, this paper makes 3 contributions: 1) replication and results confirmation of a leading technique; 2) testing straightforward modifications likely to improve performance; and 3) extension to a novel complimentary type of crisis-relevant information to demonstrate it's generalisability.

Keywords

Event-Detection, Social-Media, Crisis-Informatics, Word-Embeddings, CNN.

INTRODUCTION

The past decade has seen research using social media to automatically detect the occurrence of events of interest flourish (Atefeh and Khreich 2015; Goswami and Kumar 2016; Hasan, M. A. Orgun, et al. 2018; Cordeiro and Gama 2016; A. Weiler et al. 2017; A. Zimmermann 2014). Twitter has become the dominant platform, primarily resulting from ease of use as a broadcasting mechanism for anyone with internet connectivity, combined with ease of access to data for researchers. Huge numbers of people take to Twitter daily, publicising everything from mundane to globally important revelations. This very ease of access to, and dissemination of information renders Event Detection (ED) in social media distinctly different from traditional media. High volume and velocity throughput, coupled with hugely varied language, make Twitter a very challenging context (Zhao et al. 2011).

Despite challenges interred, potential value of effective ED in Twitter is enormous. Applications span personal social planning (Choudhury and Alani 2014; Cavalin et al. 2015), to predicting the stock market (Tsapeli et al. 2017; Alcorn 2013; Pagolu et al. 2016), among myriad others. Twitter is often the first medium where events are reported, beating even traditional news-wire (Shuai et al. 2018; Kalyanam et al. 2016; Liu et al. 2017; Thapen et al. 2016). Leveraging early reporting via accurate ED can simplify and accelerate decision making processes and provide critically decisive time advantages.

One application which has received significant attention is detection of unfolding *crises* (Imran, Castillo, Diaz, et al. 2015; Nazer et al. 2017; Said et al. 2019; Pekar et al. 2016; Snyder, Karimzadeh, et al. 2019). This includes natural disasters (e.g. earthquakes), and anthropic emergencies (e.g. terrorist attacks). For crisis-responders, such as emergency-services dispatchers, early access to pertinent information can make a huge difference to response effectiveness (Imran, Castillo, Lucas, et al. 2014; Thapen et al. 2016; Huang and Xiao 2015; Sen et al. 2015; Karami et al. 2019; Zade et al. 2018; Snyder, Karimzadeh, et al. 2019; Snyder, Lin, et al. 2019). Similarly for journalists, there's decisive advantage in early knowledge of unfolding events, where even minutes' lead can mean beating

*http://www.taglaboratory.org/
†https://www.sussex.ac.uk/

competitors to breaking stories (Liu et al. 2017; Diakopoulos et al. 2012; Repp and Ramampiaro 2018; Freitas and Ji 2016; Zubiaga 2019; Shuai et al. 2018; Hasan, M. A. Orgun, et al. 2016; Hasan, M. Orgun, et al. 2016).

Twitter data value comprises not only notification of the event occurrence, but also specific details about it, which can inform and help guide responses. Such information includes details of affected individuals/infrastructure, requests for help, and other time-sensitive information. There's been increasing research here also, aimed at providing more fine-grained classification of tweets beyond just relevance to *some* event (Tonon et al. 2017; Hu et al. 2017; Peng et al. 2019; Vargas-Calderón et al. 2019).

Despite the wide variety of techniques tested however, there's a critical problem leveraging their potential, stemming from a fundamental lack of fair and authoritative comparability between them. This stems from several interrelated factors. Foremost is the frequent unavailability of datasets used, resulting variously from: using bespoke datasets which aren't made public (e.g. Liu et al. 2017; Tonon et al. 2017; Kanojia et al. 2016; Thapen et al. 2016; Hero 2016); removal of datasets owing to institutional/copyright issues (e.g. Edinburgh-FSD corpus of Petrovic et al. 2010); erroneous artefacts such as missing IDs (e.g. Imran, S. Elbassuoni, et al. 2013; Imran, S. M. Elbassuoni, et al. 2013); insufficiently explicated processing/combining/filtering of otherwise available datasets (rendering recapitulation of specific data used impossible; e.g. Buntain et al. 2015); and so on. Furthermore, there's often variation in how ED and related information categorisation is defined, such that even two techniques using identical data can't be compared like-for-like. Finally, owing to difficulty defining the quality of outputs, there's even significant variation in evaluation criteria employed (M. Weiler A. a. G. and Scholl 2015).

Thus, parties interested in utilising ED systems have no clear means of determining SOTA for their particular application. There's often in fact, no straightforward way of delineating between techniques' efficacy, other than prominence in literature, which can be misleading about quality and merit of the research. Furthermore, researchers looking to advance capabilities have no robust, shared baseline on which to build and improve. This renders "advances" problematic, as there's no authoritative means to establish superiority between approaches. Without such, novel research has questionable value.

This paper starts addressing this insufficiency. A technique used for ED and related information types' classification is replicated and results verified. It's chosen for its demonstrated generalisability in text processing (beyond just social media), and potential to extract multiple types of pertinent information. It's carefully re-implemented, and experiments repeated, on public data. Modifications are explored, and applications extended to novel crisis-related information types. This may then serve as a foundational baseline, enabling accurate comparison with other techniques, clarifying the path to improved capabilities.

The rest of this paper is structured thus: first, the selected technique is overviewed. Second, details of data used in the original and replication are explicated. Third, replication process and experiments, including extensions, are reported. Finally, conclusions are made about the approach and problem area, informing suggestions for future work.

APPROACH

Method selection was based on:

1. recent use in research and applied context;
2. reputation for general text-processing efficacy, indicating likely ED (and related tasks) effectiveness;
3. availability of data used;
4. apparent leading/high performance based on reported results.

There's a risk the technique doesn't represent the best of extant approaches, stemming from the ambiguity which motivates this research. Given current lack of direct comparability, regardless of technique ultimately selected, there's dire need to establish a robust baseline, against which alternatives can be compared. Without such, there's no means of determining this ranking, and hence, this should be seen as a starting point upon which future research can build.

Multifarious approaches have shown merit in ED[1]. These include time series based approaches, where events are detected as frequency bursts of hashtags (Yilmaz and Hero 2016; Ozdikis, Senkul, et al. 2012), keywords (Hossny, Moschou, et al. 2018; Hossny and Mitchell 2018), and other derived features (Nützel and F. Zimmermann 2015;

[1]See the excellent surveys of Atefeh and Khreich 2015; Hasan, M. A. Orgun, et al. 2018; Imran, Castillo, Diaz, et al. 2018 for overview.

Comito et al. 2019). Similarly research has employed topic modelling (Zhu et al. 2017; B. Wang et al. 2017), clustering techniques (Alsaedi et al. 2017; Ozdikis, Karagoz, et al. 2017), semantic-web and curated knowledge bases (e.g. Tonon et al. 2017), neural networks (Burel, Saif, and Alani 2017; Kruspe 2019), among other miscellaneous and hybrid strategies (Cordeiro 2012; Fang et al. 2016). Therefore, a technique was selected with demonstrated efficacy beyond ED, in general text classification tasks, likely to perform strongly and guide subsequent evaluation and comparison.

Burel et al. (Burel, Saif, Fernandez, et al. 2017) use a simple neural network model pioneered by Yoon Kim (Kim 2014)[2]. It's designed as a generalisable sentence classification model, capable of application to various downstream tasks, and has seen widespread success, forming the basis of many subsequent CNN approaches to text classification (e.g. Wehrmann et al. 2017; Can et al. 2018; Fan et al. 2018; Chen et al. 2018; Salehinejad et al. 2017; Bian et al. 2017; Undavia et al. 2018).

Model input constitutes word embeddings representing the source text. Architecture comprises a single convolutional layer with 128 filters each of 3 sizes (3, 4, 5; i.e. 128*3=384 filters total) that extract features from the input embeddings. Convolution layers' outputs are max-pooled, before concatenation and passing through a softmax function, providing final classifications. Burel et al. report using dropout with probability 0.5 during training to mitigate over-fitting. They use ADAM gradient descent (Kingma and Ba 2014) and batch size of 256. Training is reported as 400 iterations, though early stopping to prevent over-fitting is not mentioned, nor whether optimal weights are restored. Recreating their approach, L2 regularisation was tested, though it didn't improve performance. Early stopping was also tested, and showed significant improvement. Burel et al. use 5-fold cross-validation, though specific proportions of the corpus subsets used for train/validation/test sets were not reported.

Two architecture variants were tested in their paper, with different inputs:

1. using just word embeddings derived from the tweets' texts.
2. using word embeddings and embeddings of "semantic-concepts" contained within the tweets' texts.

For word embeddings, pre-trained GoogleNews word2vec (w2v) model [3] [4] is used. For semantic-concepts' annotation, IBM's now defunct Alchemy-API is used[5]. Fortunately they make their semantic-concepts' annotations publicly available [6]. Obsolescence of Alchemy-API does mean the specific concept annotation process can't be recreated exactly for novel data (though IBM Watson contains a similar system [7], and other similar substitute alternatives are available[8]). Fortunately, as reported by Burel et al., and confirmed herein, the semantic-concepts don't significantly improve results.

In both configurations, tweets are first pre-processed to clean and tokenise text. Unfortunately very few details of the specific process are provided. Tokens are then mapped to pre-trained word embeddings, creating for each tweet a matrix of word embeddings derived from its constituent tokens. Each matrix has dimensions 300 in one axis (corresponding to the length of pre-trained embeddings' vectors), with the other being equal in length to the highest number of tokens in any of the training set tweets. Tweets with fewer tokens than this are padded to ensure consistent length. The process is illustrated in figure 1.

In the semantic-concepts augmented model, a parallel operation extracts semantic-concepts associated with named entities. An embedding space is then initialised for these, and this results in a matrix per tweet of such concepts' embeddings. Since most tweets don't contain mentions of any named entities, they're not annotated with "semantic-concepts". Hence the semantic-concepts embeddings' matrices are primarily uniform with embeddings representing "no semantic-concept". These embeddings are 30x1 vectors, justified by Burel et al as being appropriately smaller than the word embeddings, reflecting the far smaller set of concepts to be represented (i.e. far fewer semantic-concepts modelled than words). An example from Burel et al is illustrated in table 1.

Semantic-concepts embeddings are processed by an identical CNN as the architecture described above. The outputs of the text and semantic-concepts CNNs are concatenated prior to a (final) softmax layer. Figure 2 illustrates this process.

[2]There are many implementations available, e.g. https://www.kaggle.com/hamishdickson/cnn-for-sentence-classification-by-yoon-kim/data and https://paperswithcode.com/paper/convolutional-neural-networks-for-sentence

[3]https://code.google.com/archive/p/word2vec/

[4]https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

[5]https://www.ibm.com/cloud/blog/announcements/bye-bye-alchemyapi

[6]https://www.comrades-project.eu/outputs/datasets-and-ontologies/88-datasets/39-sem-crisislext26.html

[7]https://cloud.ibm.com/catalog/services/natural-language-understanding

[8]e.g. TextRazor - https://www.textrazor.com/

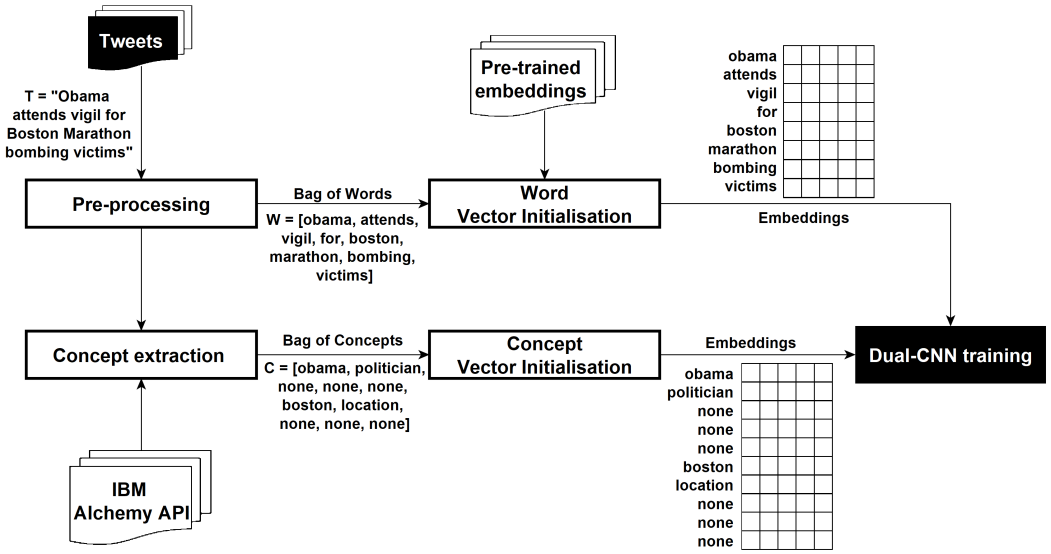

Figure 1. Pre-processing, combining word and semantic-concepts embeddings as input to Dual-CNN model. (Modified from original paper of Burel et al.)

Table 1. Semantic-Concepts annotation of Burel et al..

Original Tweet Text	'Obama attends vigil for Boston Marathon bombing victims'
Tokenised Text	['obama', 'attends', 'vigil', 'for', 'boston', 'marathon', 'bombing', 'victims']
Tokens' Semantic Concepts	['obama', 'politician', 'none', 'none', 'none', 'boston', 'location', 'none', 'none', 'none']

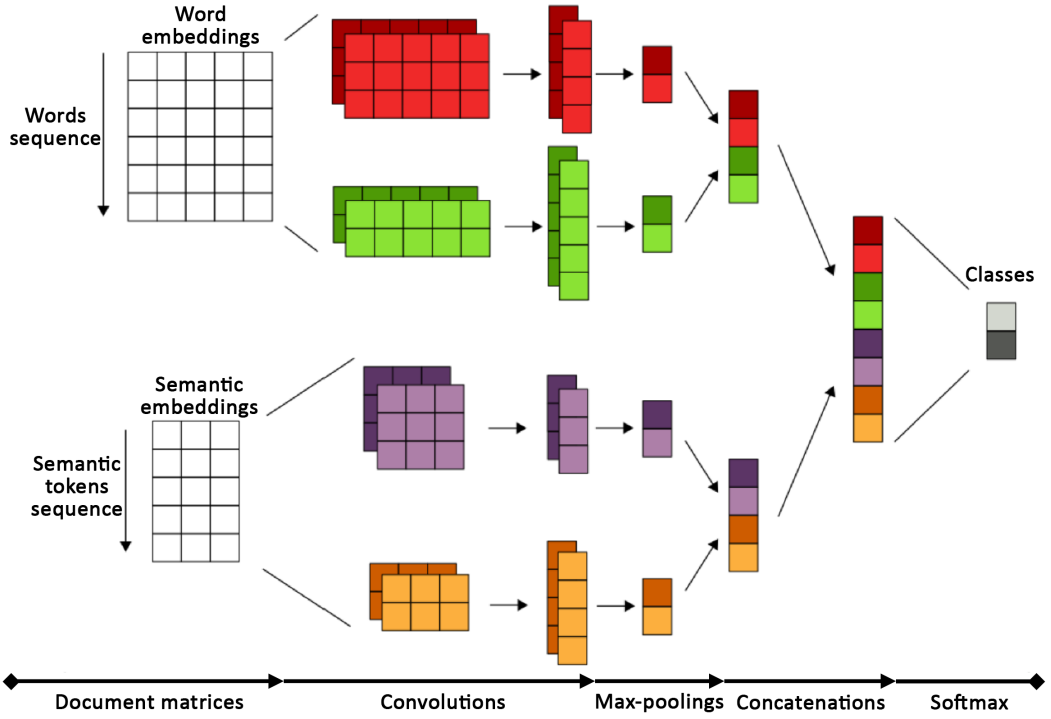

Figure 2. Dual-CNN visualisation, showing separate CNNs processing word & semantic-concepts embeddings, concatenation, and final softmax layer. (Modified from the original paper of Burel et al.)

Burel et al. justify the choice of separate CNNs by virtue of the differing length of word-based and semantic-concept based tokenisations. This precludes the more obvious introduction of the semantic-concepts as an additional channel in a (single) CNN. This however seems more a product of the tool used to perform semantic-concept annotation, which introduces additional tokens denoting the class of the named entity *in addition to the entity itself*. As seen in table 1, the annotated tokens "obama" and "boston" introduce an additional token each. Modifying the semantic-concepts' representations, to ensure length parity (e.g. replacing named entities with just their class), could enable incorporation as a separate aligned channel and potentially improve performance, but this hasn't been investigated here.

Finally, performance of the two architectures is compared to 3 baselines: Naïve Bayes (NB), Decision Trees (DT), and Support Vector Machine (SVM). Unlike the CNNs however, these aren't given the same input. Instead, they're given TF-IDF representations of the tweets' texts. It's unclear why, since this precludes direct comparison of baselines to their suggested approach, by virtue of differing representations. Ideally the baselines should also be tested with the same inputs to provide true comparison between their respective abilities to exploit those representations. This hasn't been investigated here, but remains for future work.

DATA

A primary reason for the lack of extant comparable research in ED is the dearth of high-quality annotated data. Many papers use bespoke datasets collected specifically for the purpose. Whilst valid for one-off/specific tests, the majority of these corpora aren't made public; hence performance reported in such papers becomes isolated by having no direct corollaries to compare against. No matter the works meritoriousness, it's of limited worth when it can't be directly related to others work.

Despite this, there are *some* public corpora with annotations supporting this research. Burel et al. use one such corpus, namely CrisisLexT26 (CLT26) [9], created by Olteanu, Castillo, et al. 2014 and extended by Olteanu, Vieweg, et al. 2015.

CLT26 constitutes around 26000 tweets collected between 2012-13. These comprise 26 different subsets, relating to specific individual crises, including anthropic and natural instances. The collection is annotated with 3 different attributes:

1. `Informativeness`: 4-category label indicating whether a tweet is related to the event. Categories comprise "Related and informative", "Related but not informative", "Not related" and "Not applicable". This is the target used in ED, frequently in a binary setting by collapsing the two "Related ..." labels into a single class. This is the approach adopted by Burel et al. for their first task of classifying tweets as being "related"/"unrelated" to an event.

2. `Information Type`: 7-category label indicating type of information contained in the tweet. Categories comprise: "Affected individuals", "Infrastructure & utilities", "Donations & volunteering", "Caution & advice", "Sympathy & emotional support", "Other useful info." and "Not applicable". This target can be used to train models classifying types of information contained in event-related tweets. This is used by Burel et al. in their 3rd task.

3. `Information Source`: 6-category label indicating source of the tweet. Categories comprise: "Eyewitness", "Government", "NGOs", "Business", "Media" and "Outsiders". This isn't tested in Burel et al., but is included as an extension here.

EXPERIMENTS

Burel et al. compare performance of their baselines (NB/SVM/DT) with their two CNN variants using CLT26. They test models' abilities to classify tweets for three targets, detailed in "Replications" section. I also tested several variants of their experiments, detailed in "Extensions" section.

[9] https://crisislex.org/data-collections.html#CrisisLexT26

Replications

1. *Related/Unrelated*. Binary indicator of a tweet being related to a crisis event or not. Based on "Informativeness" CLT26 annotation. Labels "Related and informative" (n=16849) and "Related-but not informative" (n=7731) are treated as positive class. Label "Not related" (n=2863) and "Not applicable" (n=489) are treated as negative class. Data is balanced by undersampling larger positive class to create final dataset of 6704 instances ((489+2863)*2).

   Curiously Burel reports final dataset size of 6703 instances; despite contacting the author, no explanatory reason is forthcoming.

2. *Event Types*. Multi-class classification of type of event tweets relate to, based on event-specific CLT26 subsets. 12 event types are defined: "shooting", "explosion", "building-collapse", "fires", "floods", "meteorite-fall", "haze", "bombing", "typhoon", "crash", "earthquake" and "derailment", based on CLT26 constituent events.

   Each Tweet from an event specific subset is of that event type; e.g. all tweets in each of the subsets "2012 Colorado wildfires", "2013 Australian bushfire" and "2013 Brazil nightclub fire" are of the "fire" event type. This same approach is replicated in this experiment. Note though that this assignment overlooks that many tweets in event specific subsets have "Informativeness" label of "Not related". Hence, they should perhaps be excluded.

   This dataset is also balanced by undersampling. Just as for the previous target though, figures reported by Burel et al. aren't consistent with expectations. i.e. The smallest event type is either Haze or Bombing, both comprising 1000 tweets. Hence, undersampling the remaining 11 categories should result in 12*1000=12000 tweets, not 12997 reported by Burel et al.. Again, the authors have been contacted, but no explanation provided.

3. *Information Types*. 7-class label indicating type of information contained in tweets. Dataset is also balanced by undersampling, and again numbers reported by Burel don't match expectations. Smallest class is "Not applicable" (1138 tweets); hence there should be 1138*7=7966 tweets, not 9105 reported by Burel et al..

Extensions

1. *Embeddings variants*. Burel et al. use the pre-trained word2vec trained on Google News. Since the style and constraints of news articles and tweets differ significantly, it should be straightforward to improve performance by using alternative pre-trained models trained on twitter data. Several are available, including the three selected:

   (a) *Godin_Twitter_400d_w2v* (Godin 2019): 400d-word2vec model trained on Twitter data[10].

   (b) *Twitter_200d_GloVe*: 200d-GloVe (Pennington et al. 2014) model trained on Twitter data [11].

   (c) *CrisisNLP_300d_w2v*: 300d-word2vec model trained on crisis-related Twitter data, from CrisisNLP (Imran, Mitra, et al. 2016) (QCRI [12])[13].

2. *Original vs re-hydrated data*. CLT26 was released in 2015 and covers data from 2012-13. Originally comprising 26000 annotated tweets, re-hydrating through Twitter's API results in reduction in size.

   Though not essential for Burel et al., any approach requiring Tweet metadata necessitates re-hydration, since the CLT26 distribution doesn't include this. Hence this information must be obtained by re-downloading Tweets from Twitter directly. Furthermore, Twitter dataset "rot" is common and often the later use of these corpora must work with a subset of the original.To gauge the impact of corpus size reduction, some experiments were repeated on re-hydrated data.

   At time of rehydration, only 69% of original tweets were still available. For individual crises, the lowest fraction was for "2013 Singapore haze", with only 54% still online. The highest fraction was for "2012 Venezuela Refinery", with 79% of Tweets available.[14]

[10] https://github.com/FredericGodin/TwitterEmbeddings
[11] https://nlp.stanford.edu/projects/glove/
[12] https://crisisnlp.qcri.org/
[13] https://crisisnlp.qcri.org/lrec2016/lrec2016.html
[14] Rehydrated-datasets, and data-configurations used for experiments available at https://github.com/j-m-crow/2020_crisis_baselines

3. *Binary/mutliclass variants*. Testing binary/multiclass equivalents of target configurations in Burel, to assess relative complexity of each:

   (a) *Multiclass Informativeness*. Burel collapses 4-class "Informativeness" to binary "Related"/"Unrelated". I include 4-class setting which provides more nuanced information could permit more effective filtering (to see i.e. just "Related and informative", rather than also including "Related but uninformative", which may impede effective use).

   (b) *Binary event types*. Burel et al. test event types in multiclass configuration. I include binary variant, splitting events into either natural/anthropic crises, indicating crises as natural disasters or the result of human actions.

   (c) *Binary information types*. Binary version of "Information types", intended to classify as either actionable (i.e. information of benefit to crisis-responders) or not (other information types which don't aid/inform crisis-response, dsepite potential relatedness). Actionable categories comprise "Affected Individuals", "Infrastructure and utilities", "Caution and advice" and "Other useful information". Non-actionable categories comprise remaining "Donations and volunteering", "Sympathy and support" and "Not applicable".[15]

4. *Information Source/Eyewitness*. The last extension adds the additional CLT26 annotation, "Information Source". Tested in both multiclass (7-class) and binary configuration, where non-"Eyewitness" labels are collapsed to singular negative class. Motivation is that often crisis-eyewitnesses can provide the most up-to-date and accurate information on situational needs. Hence, accurately identifying them could greatly aid emergency response. Indeed, there's increasing research in this area (e.g. Fang et al. 2016, Krumm and Horvitz 2015, Tanev et al. 2017, Zahra, Imran, Ostermann, et al. 2018, Morstatter et al. 2014, Diakopoulos et al. 2012, Cresci et al. 2018, Zhang et al. 2018, Truelove, Vasardani, et al. 2017, Starbird et al. 2012, Pekar et al. 2016, Snyder, Karimzadeh, et al. 2019, Truelove, Khoshelham, et al. 2017, Zahra, Imran, and Ostermann 2020).

Owing to the large number of experiment permutations comprised above and limited resource, not all configurations were tested. This permitted focussing on those most interesting from replication perspective, and providing wide baselines of the method/data to build upon.

IMPLEMENTATION

Numerous issues were encountered during implementation, concerning CLT26 and approach of Burel et al.. Hence, whilst attempting replication as accurately as possible, numerous approximations and suppositions were necessary. Since these may impact results, they're overviewed below.

CrisisLexT26

- *Overlap between labelled crises*. CLT26 comprises 26 subsets of specific-crisis related tweets. Owing to overlap in keywords used to retrieve tweets from Twitter and the time covered (all events occurring in 2012-13), there are tweets appearing in more than one subset. Since tweet annotations are relative to *individual crises*, duplicates appear with different annotations in different sub-collections. For example, Tweet-ID '354439470801616898' appears in "2013 Alberta Floods", as well as "2013 Lac Megantic train crash". "Informativeness" and "Information Source" annotations are consistent between the two, but not "Information Type". For Alberta, it's labelled "Sympathy and Support"; for Lac Megantic, it's "Other useful information".

  Such overlap is problematic, potentially meaning training set tweets appearing in validation and test sets, hence invalidating results. Additionally, lack of consistency in labels between duplicates could further obfuscate training. There's no mention of these overlaps in Burel et al., hence it's likely they weren't apparent. Fortunately the number of overlaps is small and unlikely to have *significantly* impacted results. Nevertheless, all tweets are de-duplicated in replication, to ensure discrete train/validation/test sets and valid results.

[15]Note, no single notion of "actionability" applicable to all crises and crisis-responder roles exists (Ghosh et al. 2018). Rather, actionability stems from numerous variables, including specific crisis-type, roles of responders using such information, application domain, as well as myriad other factors (Zade et al. 2018). Moreover, what comprises actionability evolves during the unfolding of crises (Munro 2011). Definition adopted here prioritises actionable information in the immediate aftermath of crisis-onset; hence excluding "donations and volunteering" which though actionable, tends to be of lower urgency (C. Wang and Lillis 2020) than other information types providing more time critical information to responders in the crucial period immediately after the crisis-onset (Mccreadie et al. 2019).

- *Malformed semantic-concepts annotations*. The semantic-concepts annotations released by Burel et al. also contain errors. Some tweets contain no entry at all (distinct from not containing annotated concepts - most CLT26 tweets contain no semantic-concepts annotations, but are recorded in a specific format). More problematically, 3 entries contain annotations with a large discrepancy between number of tweet tokens and number of tokens in the annotation. The annotation for Tweet-ID 399804790227468289 for example comprises 562 tokens (all "none") - more tokens than there are characters permitted in a tweet [16].

  Tweets with mismatched annotation lengths and missing entries are clearly errors. Since the Alchemy-API service no longer exists, it's impossible to amend these errors. Hence, tweets displaying such are dropped. Despite contacting the authors, it's unclear whether these discrepancies were apparent during experiments, or if they were introduced later, prior to public release.

Burel et al.

- *Dataset sizes discrepancies*. Outlined above, numerous discrepancies exist in dataset sizes reported in Burel et al.. Despite attempts to infer possible causation, and having contacted the authors, unfortunately no explanation has been provided.
- *Pre-processing*. Burel et al. provide scant details of pre-processing employed. Hence, I've adopted a fairly standard approach to cleaning and tokenisation of tweets that matches those few details, though the possibility remains that this diverges from the original. I pre-process tweets using the following procedure:
  1. data is cleaned (i.e. duplicate tweets and those with erroneous semantic-concepts are removed).
  2. URLs are removed using simple regex pattern-matching.
  3. non-EN unicode characters are removed.
  4. text is lowercased.
  5. tokens are extracted using Keras' Tokenizer [17] in the default configuration.
  6. tokens are mapped to pre-trained word embeddings. Out-of-vocabulary tokens are handled as detailed below. Embedding fine tuning during model training is allowed to improve efficacy.
  7. semantic-concepts are aligned with Tweets based on ID.
  8. semantic-concepts embeddings are initialised using Keras' Embedding layer. These are also fine-tuned during training to improve performance.
- *Cross-validation*. Though Burel explicitly states using 5-fold cross-validation, it's unclear how this is realised on the CLT26 data-strata. It appears folds were applied in aggregate across all subsets, which could permit data from individual crises to appear in training and testing sets. Despite inherent issues, since the aim is to validate Burel et al. results, the same strategy is adopted in this research.
- *Model parametrisation*. Initialisation and parametrisation details of CNN models and baselines aren't provided. Hence, parameter search was necessary to provide optimal results. Limited by resources though, there's potential these could be improved further.
- *Out-of-vocabulary tokens*. Burel et al. don't explicate their strategy for dealing with out-of-vocabulary tokens without entries in pre-trained word2vec model. Two common approaches are using fixed representations for all OOV tokens, or generating random embeddings for each novel token according to pre-existing embedding space distribution. Since the latter provides more granular token demarcation, this was used here.

RESULTS

Replications

Table 2 shows results of direct replications. Fortunately, results are almost uniformly confirmatory of those reported by Burel et al., lending weight to conclusions made therein. Minor exceptions exist where I was unable to match original results. These are sufficiently small to be accounted for by lack of *exhaustive* parameter search, dictated by pragmatic resource use, and not of concern. In several instances results were surpassed (e.g. Information Type 0.012 F1 increase). Likewise though, these differences are insignificant, and likely result from pragmatic training trade-offs of myself and Burel et al..

[16] https://blog.twitter.com/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html
[17] https://keras.io/preprocessing/text/

Table 2. Burel et al. results compared to replications.

(*Italics* are results of Burel et al.. Replications equalling or surpassing are marked * (those without achieve less). Each metric (*P*recision, *R*ecall, *F1*) per-target top result shown in **bold**, for both full (unbalanced) and sample (balanced) configurations.)

	Related/Unrelated						Event Types						Information Types					
Model	**P**		**R**		**F1**		**P**		**R**		**F1**		**P**		**R**		**F1**	
Full (unbalanced)																		
Naïve Bayes	*0.846*	***0.898**	*0.684*	0.633	*0.733*	0.682	*0.941*	*0.959	*0.927*	0.913	*0.933*	*0.934	*0.600*	*0.649	*0.570*	0.551	*0.579*	0.499
Decision Tree	*0.742*	*0.756	*0.707*	*0.729	*0.723*	*0.741	*0.992*	0.987	*0.992*	0.988	*0.992*	0.988	*0.506*	*0.515	*0.491*	*0.499	*0.497*	*0.506
SVM	*0.870*	*0.874	*0.738*	*0.745	*0.785*	*0.791	***0.997***	0.994	***0.996***	0.991	***0.997***	0.993	*0.642*	*0.653	*0.604*	***0.607**	*0.616*	***0.622**
CNN	*0.861*	*0.861	*0.744*	***0.762**	*0.797*	***0.800**	*0.991*	*0.995	*0.986*	*0.993	*0.988*	*0.994	*0.634*	***0.654**	*0.590*	*0.605	*0.609*	*0.621
dual-CNN	*0.857*	*0.858	***0.762***	0.751	*0.798*	0.792	*0.990*	*0.995	*0.985*	*0.994	*0.988*	*0.994	*0.648*	*0.648	*0.581*	*0.596	*0.601*	*0.613
Sample (balanced)																		
Naïve Bayes	*0.795*	*0.820	*0.787*	*0.806	*0.785*	*0.804	*0.929*	*0.934	*0.928*	*0.932	*0.928*	*0.932	*0.558*	*0.599	*0.563*	*0.566	*0.556*	*0.567
Decision Tree	*0.770*	0.767	*0.769*	0.767	*0.769*	0.767	*0.988*	0.979	*0.988*	0.978	*0.988*	0.978	*0.471*	0.466	*0.464*	0.461	*0.464*	0.462
SVM	*0.833*	*0.839	*0.830*	*0.836	*0.829*	*0.835	***0.995***	0.993	***0.995***	0.993	***0.995***	0.993	*0.606*	***0.635**	*0.609*	***0.621**	*0.605*	***0.625**
CNN	*0.839*	***0.840**	*0.838*	***0.839**	***0.838***	***0.838**	*0.983*	*0.991	*0.983*	*0.990	*0.983*	*0.991	*0.610*	*0.628	*0.610*	***0.621**	*0.610*	*0.622
dual-CNN	*0.835*	0.825	*0.833*	0.823	*0.833*	0.823	*0.985*	*0.992	*0.985*	*0.992	*0.985*	*0.992	*0.615*	*0.621	*0.615*	*0.616	*0.613*	*0.617

Concerning results variance, using a two-tailed paired t-test, measured over 5 CV runs (F1), differences between CNNs and SVM are insignificant at $\alpha = 0.05$ and $\alpha = 0.005$, for both Related/Unrelated and Information Types, in both balanced/unbalanced settings (independently). Conversely, NB and DT are shown to be significantly inferior at both α values. Additionally, differences between top performing models in each balanced/unbalanced setting are significant, at $\alpha = 0.05$ and $\alpha = 0.005$, highlighting positive effects of class-balancing this highly skewed dataset. Significance testing wasn't undertaken for Event Types target owing to likely over-fitting, detailed in Extensions results section.

Extensions

Tables 3, 4, 5 & 6 show results of extensions.

Embeddings variants

Table 3 shows results using different word-embedding models. Google News model results listed in this table are replication results, not those of Burel et al.. This ensures absolute parity and comparability between results in this table. Improvements are small but notable, with Twitter-specific Godin_w2v and GloVe providing most consistent increase. Interestingly, Crisis-tweets trained model from CrisisNLP didn't improve results as much as expected. This likely stems from specific pre-processing employed, particularly removal of non-EN characters, which pre-disposes the tweets' contents to more *standard* linguistic style, hence ensuring Google embeddings performed well despite domain mis-match.

Overall it seems so long as the embeddings are sufficiently well (pre-)trained, choice of specific model doesn't overly affect performance. Testing different pre-processing approaches would be interesting to see if these provide more distinct separation between models' efficacies, and potential to combine the multiple models as inputs, to leverage information each encodes, to see if there are notable gains. Without further experimentation, the GloVe or Godin model is the best choice.[18]

Note, Event Types wasn't tested with further embeddings (in rehydrated data configuration), since results (being so high) suggest the current task configuration is trivially easy to solve. Indeed, Burel et al. note the limited range of event types and specific event-instances in CLT26 devolves the task to essentially keyword spotting, since most crises type can be identified simply by recognising presence of certain keywords (e.g. Burel et al. note, "77% of the tweets about meteorite falls contain the word meteor"). Hence, model overfitting to these specific keyword indicators, means it would likely fail to generalise beyond this limited context. It could be possible to ameliorate this by removal of such indicators and/or increasing variety of specific event-instances. This hasn't been done here owing to limited compute-resource, and preference for testing Information Source annotation, requiring pragmatic trade-off.

[18]Note, at time of research, incorporating contextual embeddings (e.g. BERT, Devlin et al. 2019) wasn't feasible. Given widespread demonstration of such embeddings' improvements across various NLP tasks, it's expected they would be competitive with best performing embeddings here, if not superior. Whether there would be *significant* improvements though remains to be seen. In future work I plan incorporating such contextual embeddings, and combinations of different embedding sets, as noted above.

Table 3. Comparison of different embeddings for CNNs.

Each metric (*P*recision, *R*ecall, *F1*) per-target top result shown in **bold**, for full (unbalanced) and sample (balanced) configurations. Results marked * weren't tested.

			Related/Unrelated			Event Types			Information Types		
Model	**Data**	**Embeddings model**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
CNN	Full (unbalanced)	w2v_googleNews_300d	0.861	0.762	0.800	0.995	0.993	0.994	0.654	0.605	0.621
dual-CNN	Full (unbalanced)	w2v_googleNews_300d	0.858	0.751	0.792	0.995	0.994	0.994	0.648	0.596	0.613
CNN	Full (unbalanced)	godin_w2v_twitter_400d	0.851	0.765	0.800	**0.996**	0.995	0.995	0.643	0.609	0.620
dual-CNN	Full (unbalanced)	godin_w2v_twitter_400d	0.850	0.746	0.785	**0.996**	**0.996**	**0.996**	0.643	0.598	0.614
CNN	Full (unbalanced)	glove_twitter_200d	0.866	**0.776**	**0.812**	0.994	0.993	0.994	**0.655**	**0.611**	**0.624**
dual-CNN	Full (unbalanced)	glove_twitter_200d	0.867	0.760	0.801	**0.996**	0.995	0.995	0.654	0.601	0.617
CNN	Full (unbalanced)	crisisNLP_w2v_300d	**0.871**	0.760	0.802	*	*	*	*	*	*
dual-CNN	Full (unbalanced)	crisisNLP_w2v_300d	0.863	0.749	0.790	*	*	*	*	*	*
CNN	Sample (balanced)	w2v_googleNews_300d	**0.840**	**0.839**	**0.838**	0.991	0.990	0.991	0.628	0.621	0.622
dual-CNN	Sample (balanced)	w2v_googleNews_300d	0.825	0.823	0.823	0.992	0.992	0.992	0.621	0.616	0.617
CNN	Sample (balanced)	godin_w2v_twitter_400d	0.834	0.833	0.833	**0.994**	**0.994**	**0.994**	0.621	0.617	0.618
dual-CNN	Sample (balanced)	godin_w2v_twitter_400d	0.823	0.821	0.821	0.993	0.993	0.993	**0.636**	0.628	0.629
CNN	Sample (balanced)	glove_twitter_200d	0.837	0.837	0.837	0.993	0.993	0.993	0.635	**0.636**	**0.634**
dual-CNN	Sample (balanced)	glove_twitter_200d	0.835	0.834	0.834	**0.994**	**0.994**	**0.994**	0.629	0.626	0.626
CNN	Sample (balanced)	crisisNLP_w2v_300d	0.822	0.820	0.820	*	*	*	*	*	*
dual-CNN	Sample (balanced)	crisisNLP_w2v_300d	0.828	0.829	0.828	*	*	*	*	*	*

Original vs re-hydrated data

Table 4 shows results using re-hydrated data. These results provide comparison reference for methods requiring Tweet metadata which necessitate CLT26 rehydration. They also provide guidance in expected performance curtailment induced by reduction in corpus size resulting from re-hydration.

Certain metrics show increased performance (e.g. NB achieving P=0.927 on re-hydrated vs 0.898 originally), but as expected there's an decline overall. This is less extreme than expected, mostly limited to <=5% difference. Note though that this testing doesn't provide sufficient disambiguation of models true predictive capacity. To do so requires further experiments testing trained models on entirely novel data. This would likely show more marked reduction in performance owing to smaller variance in training tweets of the smaller corpus. This is left for future work.

Binary/Multiclass variants

Table 5 shows results of "Related/Unrelated" and "Information Type" in their multiclass/binary variants. Event type wasn't included owing to lack of merit (data being insufficient for reasonably representing the task).

Unsurprisingly, there's significant dropoff in efficacy moving from binary related/unrelated to 4-class relatedness configuration. F1 peaks at 0.838 for the binary case, and only 0.664 for multiclass. Whilst this doesn't render the model useless, far lower F1 does indicate likely ineffectiveness in real world deployment. Performance could be improved by increased provision of higher quality, more diverse data. Presently however, those wishing to deploy live systems, should sacrifice increased relatedness disambiguation and opt for more readily usable binary variant. Additionally, usefulness of the binary variant could be improved by reconfiguring positive class to include only "Related and informative", combining "Related but not informative" with the negative class. This would ensure any tweets classified as related are indeed informative (hence, *actionable*), and reduce noise in the deluge of data surrounding high-profile crises.

Similarly unsurprisingly, moving Information Type from binary to multiclass shows increased performance (peak multiclass F1=0.625, binary F1=0.85). This suggests crisis-responders should use the simpler, higher performing binary variant. Disambiguating actionable from non-actionable with higher precision being more useful under time pressure than noisier, less informative multiclass alternative.

Information Source/Eyewitness

Table 6 shows "Information Source" results. Various experiments were run, including embeddings variants, and using original vs re-hydrated data. This ensures parity of results provided for all CLT26 targets (which are more

Table 4. Re-hydrated data results

Each metric (*P*recision, *R*ecall, *F1*) per-target top result shown in **bold**, for full (unbalanced) and sample (balanced) configurations. Results marked * weren't tested.

			Related/Unrelated			Event Types			Information Types		
Model	**Data**	**Features**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Naïve Bayes	Full (unbalanced)	TF-IDF	**0.927**	0.606	0.651	0.958	0.898	0.925	**0.674**	0.532	0.549
Decision Tree	Full (unbalanced)	TF-IDF	0.753	0.714	0.731	0.983	0.986	0.985	0.505	0.485	0.493
SVM	Full (unbalanced)	TF-IDF	0.876	0.737	**0.786**	**0.994**	**0.992**	**0.993**	0.659	**0.589**	**0.612**
CNN	Full (unbalanced)	w2v_googleNews_300d	0.844	0.713	0.758	*	*	*	*	*	*
dual-CNN	Full (unbalanced)	w2v_googleNews_300d	0.848	0.715	0.760	*	*	*	*	*	*
CNN	Full (unbalanced)	godin_w2v_twitter_400d	0.824	0.711	0.752	*	*	*	*	*	*
dual-CNN	Full (unbalanced)	godin_w2v_twitter_400d	0.834	0.72	0.761	*	*	*	*	*	*
CNN	Full (unbalanced)	glove_twitter_200d	0.861	0.735	0.781	*	*	*	*	*	*
dual-CNN	Full (unbalanced)	glove_twitter_200d	0.851	**0.739**	0.781	*	*	*	*	*	*
CNN	Full (unbalanced)	crisisNLP_w2v_300d	0.855	0.713	0.759	*	*	*	*	*	*
dual-CNN	Full (unbalanced)	crisisNLP_w2v_300d	0.848	0.729	0.770	*	*	*	*	*	*
Naïve Bayes	Sample (balanced)	TF-IDF	0.794	0.776	0.772	0.941	0.936	0.936	0.586	0.579	0.570
Decision Tree	Sample (balanced)	TF-IDF	0.747	0.746	0.746	0.976	0.975	0.975	0.462	0.452	0.454
SVM	Sample (balanced)	TF-IDF	0.815	**0.810**	**0.809**	**0.990**	**0.990**	**0.990**	**0.610**	**0.593**	**0.597**
CNN	Sample (balanced)	w2v_googleNews_300d	0.810	0.805	0.804	*	*	*	*	*	*
dual-CNN	Sample (balanced)	w2v_googleNews_300d	0.805	0.802	0.801	*	*	*	*	*	*
CNN	Sample (balanced)	godin_w2v_twitter_400d	**0.816**	0.809	**0.809**	*	*	*	*	*	*
dual-CNN	Sample (balanced)	godin_w2v_twitter_400d	0.812	0.808	0.808	*	*	*	*	*	*
CNN	Sample (balanced)	glove_twitter_200d	0.802	0.799	0.799	*	*	*	*	*	*
dual-CNN	Sample (balanced)	glove_twitter_200d	0.804	0.802	0.802	*	*	*	*	*	*
CNN	Sample (balanced)	crisisNLP_w2v_300d	0.791	0.790	0.790	*	*	*	*	*	*
dual-CNN	Sample (balanced)	crisisNLP_w2v_300d	0.802	0.798	0.798	*	*	*	*	*	*

generally four commonly specified targets in crisis-response). Furthermore, there's increasing interest in identifying information sources, since it provides incisive means of filtering those (few) sources most likely to be able to provide timely and pertinent information during crises. Particularly interesting is binary configuration of being an Eyewitness or not (i.e. any of the other classes). Eyewitnesses, co-located at the geographical location of events' occurrence by definition, are best placed to provide accurate, up to date information.

Results are promising, and comparable to model performance on other CLT26 targets. Embedding model choice again has relatively minor impact. Binary configuration shows far higher efficacy than multiclass. Indeed, binary configuration F1 suggests real-world usefulness, whilst multiclass model (peak F1=0.499) doesn't demonstrate sufficient efficacy to be useful.

Interestingly, peak performance is by SVM using balanced (original) data, with numerous possible explanations. Neural nets, requiring significantly greater volumes of training data, may be disadvantaged by virtue of limited size of CLT26, and additional class-balancing size dropoff. Additionally, CNNs spatial aspect may not provide sufficient advantage over SVM in the context of short tweets. Relatedly, specific Eyewitness tweets' syntax may be such that spatial language characteristics that can be leveraged by CNNs are simply not present. Further work is required to disambiguate this issue.

CONCLUSIONS & FUTURE WORK

This paper demonstrates numerous interesting aspects of crisis-information classification on Twitter. It's also highlighted broader key issues in reproducibility of machine-learning and social media focussed research. Foremost, it's validated the baselines of Burel et al. through non-trivial replication. Moreover, various extensions have demonstrated generalisability of the CNN design of Kim, and wide ranging efficacy in short-text classification. Novel formulations of additional targets and test configurations are of significant interest and practical use to researchers and practitioners harnessing such systems, providing nuanced guidance for appraisal and selection for real-world deployment.

Table 5. Binary/Multiclass variants results

Each metric (*P*recision, *R*ecall, *F1*) per-target top result shown in **bold**, for full (unbalanced) and sample (balanced) configurations.

			Related/Unrelated (Multiclass)			Information Types (binary)		
Model	**Data**	**Features**	**P**	**R**	**F1**	**P**	**R**	**F1**
Naïve Bayes	Full (unbalanced)	TF-IDF	**0.755**	0.487	0.521	0.858	0.833	0.843
Decision Tree	Full (unbalanced)	TF-IDF	0.576	0.521	0.542	0.780	0.775	0.777
SVM	Full (unbalanced)	TF-IDF	0.722	**0.597**	**0.632**	**0.859**	**0.837**	**0.846**
CNN	Full (unbalanced)	w2v_googleNews_300d	0.747	0.578	0.610	0.841	0.836	0.838
dual-CNN	Full (unbalanced)	w2v_googleNews_300d	0.736	0.572	0.604	0.847	0.835	0.840
Naïve Bayes	Sample (balanced)	TF-IDF	0.634	0.621	0.612	0.846	0.845	0.845
Decision Tree	Sample (balanced)	TF-IDF	0.450	0.449	0.448	0.758	0.758	0.758
SVM	Sample (balanced)	TF-IDF	**0.667**	**0.666**	**0.664**	**0.850**	**0.850**	**0.850**
CNN	Sample (balanced)	w2v_googleNews_300d	0.610	0.598	0.599	0.849	0.848	0.848
dual-CNN	Sample (balanced)	w2v_googleNews_300d	0.577	0.566	0.565	0.849	0.848	0.848

The need for robustifying approaches to continued research in this area was also highlighted. Numerous meritorious research avenues are being pursued in this and related domains. However, without *authoritative* and *comparable* baselines against which to measure these, there's no means to quanitfy the significance of this "progress". This paper provides one such baseline, enabling better understanding of SOTA and a yardstick for comparison of improvements.

Additionally, several future work trajectories were identified. Directly, this comprises extension to more diverse extant modelling approaches. It also includes potential to improve upon the effective, simple and generalisable CNN model, in both straightforward and more complex ways.

Attaining the above contributions also importantly highlighted discrepancies in reporting of corpus statistics, and uncovered numerous errors in modelling approach and corpus annotations. This has relevance beyond ED, in the broader challenge of reproducibility in machine learning research. Indeed, primary difficulties conducting this research stemmed from lack of thorough and exhaustive process reporting, and frequent obfuscation by omission in data and modelling reporting, of crucial factors affecting feasibility of independently replication.

Paramount importance of researchers reporting details and implications of research as completely as possible can't be overstated. It's imperative that either research code be shared, or data preparation, pre-processing methodologies, model parametrisation, and experimental context be fully explained to enable independent recreation thereof. Similarly, where data issues arise, either during creation or utilisation, these must be reported clearly and cogently, whilst examining potential impacts, in order to maintain research validity. Public data sources must be utilised, either independently curated and made available, or through structured challenges such as TREC incident streams[19]. Where potential for errors/ambiguities persists, these must be openly reported to enable their continued tackling and resolution by the wider community.

Significant time was expended ameliorating these issues, and I hope the demonstration of such detailed and thorough inspection required invigorates the community to focus more effort here. Whilst it may not have the headline appeal of superficial advances on the SOTA, ensuring the field is supported by robust, verifiable methodologies at its foundation, is arguably of far more importance.

Finally, the need for more nuanced error analyses in this and similar work was emphasised, alongside the dire need for higher quality, larger and more diverse corpora. Continuing, it's my intention to pursue both axes - increasing quality and incisiveness of crisis-classification error-analyses, parallel to improving data provision. In so doing I hope a collection of resources can be created and made available, directly building upon those already available from this research.

[19] http://dcs.gla.ac.uk/~richardm/TREC_IS/

Table 6. Information Source results

Each metric (*P*recision, *R*ecall, *F1*) top result shown in **bold**, for full (unbalanced) and sample (balanced) configurations, per original and re-hydrated data. Results marked * weren't tested.

			Binary			Multiclass		
Model	**Data**	**Features**	**P**	**R**	**F1**	**P**	**R**	**F1**
Naïve Bayes	Original / full (unbalanced)	TF-IDF	**0.856**	0.626	0.675	**0.643**	0.327	0.365
Decision Tree	Original / Full (unbalanced)	TF-IDF	0.720	0.702	0.711	0.407	0.360	0.377
SVM	Original / Full (unbalanced)	TF-IDF	0.831	0.724	0.764	0.601	**0.459**	**0.499**
CNN	Original / Full (unbalanced)	w2v_googleNews_300d	0.823	0.733	0.767	0.579	0.444	0.482
dual-CNN	Original / Full (unbalanced)	w2v_googleNews_300d	0.815	0.728	0.761	0.584	0.442	0.476
CNN	Original / Full (unbalanced)	godin_w2v_twitter_400d	0.793	0.725	0.753	0.589	0.448	0.478
dual-CNN	Original / Full (unbalanced)	godin_w2v_twitter_400d	0.799	0.709	0.744	0.600	0.450	0.479
CNN	Original / Full (unbalanced)	glove_twitter_200d	0.833	**0.737**	**0.774**	0.575	0.458	0.493
dual-CNN	Original / Full (unbalanced)	glove_twitter_200d	0.826	0.735	0.771	0.578	0.457	0.492
CNN	Original / Full (unbalanced)	crisisNLP_w2v_300d	0.810	**0.737**	0.767	*	*	*
dual-CNN	Original / Full (unbalanced)	crisisNLP_w2v_300d	0.814	0.711	0.749	*	*	*
Naïve Bayes	Original / Sample (balanced)	TF-IDF	0.806	0.806	0.806	0.479	0.466	0.463
Decision Tree	Original / Sample (balanced)	TF-IDF	0.732	0.731	0.731	0.316	0.313	0.312
SVM	Original / Sample (balanced)	TF-IDF	**0.821**	**0.820**	**0.820**	0.479	0.473	**0.472**
CNN	Original / Sample (balanced)	w2v_googleNews_300d	0.809	0.806	0.806	**0.481**	0.467	0.469
dual-CNN	Original / Sample (balanced)	w2v_googleNews_300d	0.804	0.804	0.804	0.455	0.436	0.439
CNN	Original / Sample (balanced)	godin_w2v_twitter_400d	0.808	0.806	0.806	0.485	0.470	0.470
dual-CNN	Original / Sample (balanced)	godin_w2v_twitter_400d	0.798	0.797	0.797	0.472	0.457	0.457
CNN	Original / Sample (balanced)	glove_twitter_200d	0.817	0.816	0.816	0.476	**0.484**	0.475
dual-CNN	Original / Sample (balanced)	glove_twitter_200d	0.808	0.807	0.807	0.465	0.455	0.457
CNN	Original / Sample (balanced)	crisisNLP_w2v_300d	0.810	0.809	0.809	*	*	*
dual-CNN	Original / Sample (balanced)	crisisNLP_w2v_300d	0.807	0.805	0.805	*	*	*
Naïve Bayes	Re-hydrated / full (unbalanced)	TF-IDF	**0.835**	0.560	0.585	**0.608**	0.295	0.327
Decision Tree	Re-hydrated / Full (unbalanced)	TF-IDF	0.662	0.644	0.652	0.381	0.345	0.358
SVM	Re-hydrated / Full (unbalanced)	TF-IDF	0.810	0.670	0.715	0.558	**0.414**	**0.452**
CNN	Re-hydrated / Full (unbalanced)	w2v_googleNews_300d	0.794	0.681	0.721	*	*	*
dual-CNN	Re-hydrated / Full (unbalanced)	w2v_googleNews_300d	0.776	0.680	0.714	*	*	*
CNN	Re-hydrated / Full (unbalanced)	godin_w2v_twitter_400d	0.760	0.676	0.707	*	*	*
dual-CNN	Re-hydrated / Full (unbalanced)	godin_w2v_twitter_400d	0.757	0.665	0.698	*	*	*
CNN	Re-hydrated / Full (unbalanced)	glove_twitter_200d	0.811	**0.703**	**0.743**	*	*	*
dual-CNN	Re-hydrated / Full (unbalanced)	glove_twitter_200d	0.796	0.692	0.729	*	*	*
CNN	Re-hydrated / Full (unbalanced)	crisisNLP_w2v_300d	0.781	0.665	0.701	*	*	*
dual-CNN	Re-hydrated / Full (unbalanced)	crisisNLP_w2v_300d	0.789	0.654	0.695	*	*	*
Naïve Bayes	Re-hydrated / Sample (balanced)	TF-IDF	0.806	0.805	0.805	**0.466**	0.443	0.444
Decision Tree	Re-hydrated / Sample (balanced)	TF-IDF	0.727	0.727	0.726	0.319	0.310	0.311
SVM	Re-hydrated / Sample (balanced)	TF-IDF	**0.814**	**0.813**	**0.813**	0.460	**0.454**	**0.454**
CNN	Re-hydrated / Sample (balanced)	w2v_googleNews_300d	0.779	0.778	0.778	*	*	*
dual-CNN	Re-hydrated / Sample (balanced)	w2v_googleNews_300d	**0.814**	**0.813**	**0.813**	*	*	*
CNN	Re-hydrated / Sample (balanced)	godin_w2v_twitter_400d	0.795	0.794	0.794	*	*	*
dual-CNN	Re-hydrated / Sample (balanced)	godin_w2v_twitter_400d	0.787	0.786	0.786	*	*	*
CNN	Re-hydrated / Sample (balanced)	glove_twitter_200d	0.791	0.791	0.791	*	*	*
dual-CNN	Re-hydrated / Sample (balanced)	glove_twitter_200d	0.794	0.794	0.793	*	*	*
CNN	Re-hydrated / Sample (balanced)	crisisNLP_w2v_300d	0.797	0.797	0.797	*	*	*
dual-CNN	Re-hydrated / Sample (balanced)	crisisNLP_w2v_300d	0.793	0.790	0.789	*	*	*

REFERENCES

Alcorn, S. (Apr. 2013). *Twitter Can Predict The Stock Market, If You're Reading The Right Tweets*. URL: https://www.fastcompany.com/2681873/twitter-can-predict-the-stock-market-if-youre-reading-the-right-tweets (visited on 06/14/2017).

Alsaedi, N., Burnap, P., and Rana, O. (2017). "Sensing Real-World Events Using Social Media Data and a Classification-Clustering Framework". In: *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 216–223.

Atefeh, F. and Khreich, W. (Feb. 2015). "A Survey of Techniques for Event Detection in Twitter". In: *Computational Intelligence* 31, pp. 132–164.

Bian, W., Li, S., Yang, Z., Chen, G., and Lin, Z. (2017). "A Compare-Aggregate Model with Dynamic-Clip Attention for Answer Selection". In: *ACM on Conference on Information and Knowledge Management*, pp. 1987–1990.

Buntain, C., Lin, J., and Golbeck, J. (2015). "Learning to Discover Key Moments in Social Media Streams." In: *CoRR* abs/1508.00488.

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.

Burel, G., Saif, H., Fernandez, M., and Alani, H. (2017). "On semantics and deep learning for event detection in crisis situations". In: *Workshop on Semantic Deep Learning at ESWC*.

Can, D.-C., Ho, T.-N., and Siong, C. E. (2018). "A Hybrid Deep Learning Architecture for Sentence Unit Detection". In: *International Conference on Asian Language Processing*, pp. 129–132.

Cavalin, P., G. Moyano, L., and P. Miranda, P. (Nov. 2015). "A Multiple Classifier System for Classifying Life Events on Social Media". In: *IEEE International Conference on Data Mining Workshop*, pp. 1332–1335.

Chen, S., Peng, C., Cai, L., and Guo, L. (2018). "A Deep Neural Network Model for Target-based Sentiment Analysis". In: *IJCNN*, pp. 1–7.

Choudhury, S. and Alani, H. (2014). "Personal life event detection from social media". In: *Social Personalisation Workshop at ACM Hypertext and Social Media Conference*.

Comito, C., Forestiero, A., and Pizzuti, C. (Aug. 2019). "Bursty Event Detection in Twitter Streams". In: *ACM Transactions on Knowledge Discovery from Data* 13.13, pp. 1–28.

Cordeiro, M. (2012). "Twitter event detection: combining wavelet analysis and topic inference summarization". In: *7th Doctoral symposium in informatics engineering*, pp. 11–16.

Cordeiro, M. and Gama, J. (2016). "Online Social Networks Event Detection: A Survey". In: *Solving Large Scale Learning Tasks. Challenges and Algorithms*. Lecture Notes in Computer Science. Springer, pp. 1–41.

Cresci, S., Cimino, A., Avvenuti, M., Tesconi, M., and Dell'Orletta, F. (2018). "Real-World Witness Detection in Social Media via Hybrid Crowdsensing". In: *Twelfth International AAAI Conference on Web and Social Media*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Annual Conference of the North American Chapter of the ACL: Human Language Technologies*.

Diakopoulos, N., De Choudhury, M., and Naaman, M. (2012). "Finding and assessing social media information sources in the context of journalism". In: *ACM annual conference on Human Factors in Computing Systems*. ACM.

Fan, M., Lin, W., Feng, Y., Sun, M., and Li, P. (2018). "A Globalization-Semantic Matching Neural Network for Paraphrase Identification". In: *27th ACM International Conference on Information and Knowledge Management*, pp. 2067–2075.

Fang, R., Nourbakhsh, A., LIU, X., Shah, S., and Li, Q. (2016). "Witness Identification in Twitter". In: *Fourth International Workshop on Natural Language Processing for Social Media*. ACL, pp. 65–73.

Freitas, J. and Ji, H. (2016). "Identifying News from Tweets". In: *First Workshop on NLP and Computational Social Science (NLP+CSS@EMNLP)*, pp. 11–16.

Ghosh, S., Ghosh, K., Ganguly, D., Chakraborty, T., Jones, G. J. F., Moens, M.-F., and Imran, M. (Oct. 2018). "Exploitation of Social Media for Emergency Relief and Preparedness: Recent Research and Trends". In: *Information Systems Frontiers* 20.5, pp. 901–907.

Godin, F. (2019). "Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing". PhD thesis. Ghent University, Belgium.

Goswami, A. and Kumar, A. (2016). "A survey of event detection techniques in online social networks". In: *Social Network Analysis and Mining* 6.1, pp. 1–25.

Hasan, M., Orgun, M., and Schwitter, R. (2016). "TwitterNews+: A framework for real time event detection from the twitter data stream". In: *Lecture Notes in Computer Science* 10046 LNCS, pp. 224–239.

Hasan, M., Orgun, M. A., and Schwitter, R. (2016). "TwitterNews: real time event detection from the Twitter data stream". In: *PeerJ PrePrints* 4, e2297v1.

Hasan, M., Orgun, M. A., and Schwitter, R. (Mar. 2018). "A survey on real-time event detection from the Twitter data stream". In: *Journal of Information Science* 44.4, pp. 443–463.

Hero, A. (Jan. 2016). "Multimodal Event Detection in Twitter Hashtag Networks". In: *Journal of Signal Processing Systems*.

Hossny, A. H. and Mitchell, L. (Jan. 2018). "Event detection in Twitter: A keyword volume approach". In: *IEEE International Conference on Data Mining Workshops*.

Hossny, A. H., Moschou, T., Osborne, G., Mitchell, L., and Lothian, N. (July 2018). "Enhancing keyword correlation for event detection in social networks using SVD and k-means: Twitter case study". In: *Social Network Analysis and Mining* 8.

Hu, Z., Rahimtoroghi, E., and Walker, M. (2017). "Inference of Fine-Grained Event Causality from Blogs and Films". In: *ACL Events and Stories in the News Workshop*, pp. 52–58.

Huang, Q. and Xiao, Y. (Aug. 2015). "Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery". In: *ISPRS International Journal of Geo-Information* 4.3, pp. 1549–1568.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (June 2015). "Processing Social Media Messages in Mass Emergency: A Survey". In: *ACM Comput. Surv.* 47.4, 67:1–67:38.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2018). "Processing Social Media Messages in Mass Emergency: Survey Summary". In: *Companion of The Web Conference '18*, pp. 507–511.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: Artificial Intelligence for Disaster Response". In: *23rd International Conference on World Wide Web*. ACM, pp. 159–162.

Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F., and Meier, P. (2013). "Extracting information nuggets from disaster-related messages in social media". In: *10th International ISCRAM Conference*.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). "Practical extraction of disaster-relevant information from social media". In: *22nd international conference on World Wide Web companion*, pp. 1021–1024.

Imran, M., Mitra, P., and Castillo, C. (2016). "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *10th Language Resources and Evaluation Conference*.

Kalyanam, J., Quezada, M., Poblete, B., and Lanckriet, G. (Dec. 2016). "Prediction and Characterization of High-Activity Events in Social Media Triggered by Real-World News". In: *PLOS ONE* 11.12.

Kanojia, D., Kumar, V., and Ramamritham, K. (Sept. 2016). "Civique: Using Social Media to Detect Urban Emergencies". In: *Very Large Databases 2016*.

Karami, A., Shah, V., Vaezi, R., and Bansal, A. (2019). "Twitter speaks: A case of national disaster situational awareness". In: *Journal of Information Science*. eprint: https://doi.org/10.1177/0165551519828620.

Kim, Y. (2014). "Convolutional Neural Networks for Sentence Classification". In: *ACL Conference on Empirical Methods in Natural Language Processing*, pp. 1746–1751.

Kingma, D. P. and Ba, J. (2014). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations*.

Krumm, J. and Horvitz, E. (2015). "Eyewitness: Identifying Local Events via Space-Time Signals in Twitter Feeds". In: *23rd SIGSPATIAL International Conference*.

Kruspe, A. (Oct. 2019). "Few-shot tweet detection in emerging disaster events". In: *AI+HADR Workshop @ NeurIPS*.

Liu, X., Nourbakhsh, A., Li, Q., Shah, S., Martin, R., and Duprey, J. (Nov. 2017). "Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data". In: *IEEE International Conference on Big Data*, pp. 1483–1493.

Mccreadie, R., Buntain, C., and Soboroff, I. (2019). "TREC Incident Streams: Finding Actionable Information on Social Media". In: *16th International ISCRAM Conference*.

Morstatter, F., Lubold, N., Pon-Barry, H., Pfeffer, J., and Liu, H. (2014). "Finding Eyewitness Tweets During Crises". In: *ACL Workshop on Language Technologies and Computational Social Science*, pp. 23–27.

Munro, R. (2011). "Subword and Spatiotemporal Models for Identifying Actionable Information in Haitian Kreyol". In: *Fifteenth Conference on Computational Natural Language Learning*, pp. 68–77.

Nazer, T. H., Xue, G., Ji, Y., and Liu, H. (2017). "Intelligent Disaster Response via Social Media Analysis A Survey". In: *SIGKDD Explor. Newsl.* 19.1, pp. 46–59.

Nützel, J. and Zimmermann, F. (Aug. 2015). "Improved Burst Based Real-Time Event Detection Using Adaptive Reference Corpora". In: *3rd International Conference on Future Internet of Things and Cloud*, pp. 512–518.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises". In: *8th International AAAI Conference on Web and Social Media*.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to Expect When the Unexpected Happens: Social Media Communications Across Crises". In: *18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 994–1009.

Ozdikis, O., Karagoz, P., and Oğuztüzün, H. (Dec. 2017). "Incremental clustering with vector expansion for online event detection in microblogs". In: *Social Network Analysis and Mining* 7.1, p. 56.

Ozdikis, O., Senkul, P., and Oguztuzun, H. (2012). "Semantic Expansion of Hashtags for Enhanced Event Detection in Twitter". In: *VLDB Workshop on Online Social Systems*.

Pagolu, V. S., Challa, K. N. R., Panda, G., and Majhi, B. (2016). "Sentiment analysis of Twitter data for predicting stock market movements". In: *International Conference on Signal Processing, Communication, Power and Embedded System*, pp. 1345–1350.

Pekar, V., Binner, J., Najafi, H., and Hale, C. (2016). "Selecting Classification Features for Detection of Mass Emergency Events on Social Media". In: *15th Annual International Conference on Security and Management*.

Peng, H., Li, J., Gong, Q., Song, Y., Ning, Y., Lai, K., and Yu, P. S. (2019). "Fine-grained Event Categorization with Heterogeneous Graph Convolutional Networks". In: *Twenty-Eighth International Joint Conference on Artificial Intelligence*. ijcai.org, pp. 3238–3245.

Pennington, J., Socher, R., and Manning, C. (2014). "Glove: Global Vectors for Word Representation". In: *Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1532–1543.

Petrovic, S., Osborne, M., and Lavrenko, V. (2010). "Streaming first story detection with application to Twitter". In: *Human Language Technologies - Annual Conference of the North American Chapter of the ACL*.

Repp, Ø. and Ramampiaro, H. (July 2018). "Extracting News Events from Microblogs". In: *Journal of Statistics and Management Systems* 21.4.

Said, N., Ahmad, K., Regular, M., Pogorelov, K., Hasan, L., Ahmad, N., and Conci, N. (2019). "Natural disasters detection in social media and satellite imagery: a survey". In: *Multimedia Tools and Applications* 78, pp. 31267–31302.

Salehinejad, H., Barfett, J., Aarabi, P., Valaee, S., Colak, E., Gray, B., and Dowdell, T. (2017). "A convolutional neural network for search term detection". In: *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications*.

Sen, A., Rudra, K., and Ghosh, S. (2015). "Extracting situational awareness from microblogs during disaster events". In: *7th International Conference on Communication Systems and Networks*.

Shuai, X., Liu, X., Nourbakhsh, A., Shah, S., and Custis, T. (2018). "TipMaster: A Knowledge Base of Authoritative Local News Sources on Social Media". In: *Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence*.

Snyder, L. S., Karimzadeh, M., Stober, C., and Ebert, D. S. (2019). "Situational Awareness Enhanced through Social Media Analytics: A Survey of First Responders". In: *IEEE International Symposium on Technologies for Homeland Security*.

Snyder, L. S., Lin, Y.-S., Karimzadeh, M., Goldwasser, D., and Ebert, D. S. (2019). "Interactive Learning for Identifying Relevant Tweets to Support Real-time Situational Awareness". In: *IEEE Transactions on Visualization and Computer Graphics* 26.1.

Starbird, K., Muzny, G., and Palen, L. (2012). "Learning from the Crowd: Collaborative Filtering Techniques for Identifying On-the-Ground Twitterers during Mass Disruptions". In: *9th International ISCRAM Conference*.

Tanev, H., Zavarella, V., and Steinberger, J. (2017). "Monitoring disaster impact: detecting micro-events and eyewitness reports in mainstream and social media". In: *14th International ISCRAM Conference*.

Thapen, N. A., Simmie, D. S., and Hankin, C. (2016). "The early bird catches the term: combining twitter and news data for event detection and situational awareness". In: *J. Biomedical Semantics* 7, p. 61.

Tonon, A., Cudré-Mauroux, P., Blarer, A., Lenders, V., and Motik, B. (2017). "ArmaTweet: Detecting events by semantic tweet analysis". In: *European Semantic Web Conference*, pp. 138–153.

Truelove, M., Khoshelham, K., McLean, S., Winter, S., and Vasardani, M. (Apr. 2017). "Identifying Witness Accounts from Social Media Using Imagery". In: *ISPRS International Journal of Geo-Information* 6.4.

Truelove, M., Vasardani, M., and Winter, S. (Dec. 2017). "Testing the event witnessing status of micro-bloggers from evidence in their micro-blogs". In: *PLOS ONE* 12.12.

Tsapeli, F., Bezirgiannidis, N., Tino, P., and Musolesi, M. (June 2017). "Linking Twitter Events With Stock Market Jitters". In: *CoRR* abs/1709.06519. arXiv: 1709.06519.

Undavia, S., Meyers, A., and Ortega, J. (2018). "A Comparative Study of Classifying Legal Documents with Neural Networks". In: *FedCSIS*, pp. 515–522.

Vargas-Calderón, V., Parra-A., N., Camargo, J. E., and Vinck-Posada, H. (Nov. 2019). "Event detection in Colombian security Twitter news using fine-grained latent topic analysis". In: *CoRR* abs/1911.08370. arXiv: 1911.08370.

Wang, B., Liakata, M., Zubiaga, A., and Procter, R. (Sept. 2017). "A Hierarchical Topic Modelling Approach for Tweet Clustering". In: *9th International Conference on Social Informatics*, pp. 378–390.

Wang, C. and Lillis, D. (2020). "Classification for Crisis-Related Tweets Leveraging Word Embeddings and Data Augmentation". In: *Twenty-Eighth Text Retrieval Conference*.

Wehrmann, J., Becker, W., Cagnini, H. E. L., and Barros, R. C. (2017). "A character-based convolutional neural network for language-agnostic Twitter sentiment analysis". In: *International Joint Conference on Neural Networks*, pp. 2384–2391.

Weiler, A., Grossniklaus, M., and Scholl, M. H. (Mar. 2017). "Survey and Experimental Analysis of Event Detection Techniques for Twitter". In: *The Computer Journal* 60.3, pp. 329–346.

Weiler Andreas and Grossniklaus, M. and Scholl, M. H. (2015). "Evaluation Measures for Event Detection Techniques on Twitter Data Streams". In: *Data Science*. Springer, pp. 108–119.

Yilmaz, Y. and Hero, A. (Jan. 2016). "Multimodal Event Detection in Twitter Hashtag Networks". In: *Journal of Signal Processing Systems* 90.2.

Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M., and Starbird, K. (2018). "From Situational Awareness to Actionability: Towards Improving the Utility of Social Media Data for Crisis Response". In: *ACM Conference on Human-Computer Interaction*.

Zahra, K., Imran, M., and Ostermann, F. O. (2020). "Automatic identification of eyewitness messages on twitter during disasters". In: *Information Processing & Management* 57.1, p. 102107.

Zahra, K., Imran, M., Ostermann, F. O., Boersma, K., and Tomaszewski, B. (2018). "Understanding eyewitness reports on Twitter during disasters". In: *15th International ISCRAM Conference*.

Zhang, H., Ma, F., Li, Y., Zhang, C., Wang, T., Wang, Y., Gao, J., and Su, L. (Aug. 2018). "Leveraging the Power of Informative Users for Local Event Detection". In: *IEEE/ACM International ASONAM Conference*, pp. 429–436.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). "Comparing Twitter and Traditional Media Using Topic Models". In: *Advances in Information Retrieval*. Lecture Notes in Computer Science. Springer, pp. 338–349.

Zhu, R., Zhang, A., Peng, J., and Zhai, C. (2017). "Exploiting temporal divergence of topic distributions for event detection". In: *2016 IEEE International Conference on Big Data*, pp. 164–171.

Zimmermann, A. (2014). "On the cutting edge of event detection from social streams–a non-exhaustive survey". In: *SemanticScholar*.

Zubiaga, A. (2019). “Mining Social Media for Newsgathering: A Review”. In: *Online Social Networks and Media* 13.