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 and in the aftermath of crisis events such as earthquakes and terrorist attacks. Despite myriad challenges, with the right tools, significant insight can be gained which can assist emergency responders and related applications. However, most extant approaches are incomparable, using bespoke definitions, models, datasets and even evaluation metrics. Furthermore, it is rare that code, trained models, or exhaustive parametrisation details are made openly available. Thus, even confirmation of self-reported performance is problematic; authoritatively determining the state of the art (SOTA) is essentially impossible. Consequently, to begin addressing such endemic ambiguity, this paper seeks to make 3 contributions: 1) the replication and results confirmation of a leading (and generalisable) technique; 2) testing straightforward modifications of the technique likely to improve performance; and 3) the extension of the technique to a novel and 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 a flourishing of research using social media to automatically detect the occurrence of events of interest (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 for such systems, primarily resulting from the ease of use as a broadcasting mechanism for any individual with internet connectivity, combined with the parallel ease of access to the data for researchers. Huge numbers of people take to Twitter on a daily basis, publicising everything from mundane and frivolous personal anecdotes to globally important revelations. This very ease of access to, and dissemination of information renders the challenge of Event Detection (ED) in social media distinctly different from traditional media. High throughput, in terms of both velocity and volume, coupled with hugely varied language individuals use to express themselves, make Twitter a very challenging context for this task (Zhao et al. 2011).

Despite the challenges interred, the potential value of effective ED in Twitter is enormous. Applications range from personal social planning and alerts (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. For a number of applications, Twitter is often the first medium through which events are reported, beating even traditional news-wire (Shuai et al. 2018; Kalyanam et al. 2016; Liu et al. 2017; Thapen et al. 2016). Being able to leverage this early reporting via accurate detection of events can simplify and accelerate decision making processes or provide a critical and decisive time advantage.

One application which has received significant research attention is the detection of events relating to 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). Such events include both natural disasters such as earthquakes and floods, and anthropic emergencies such as vehicular accidents, violent crime and terrorist attacks. For crisis responders, such as those managing the dispatch of emergency services and coordinating volunteer efforts, early access to pertinent information can make a

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

huge difference to the effectiveness of their response efforts (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 is a decisive advantage in having early access to knowledge of unfolding events, where even a few minutes' lead can mean beating competitors to breaking a story (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).

In almost all scenarios, whether crisis-response coordination, journalism, or other applications not explicated here, the value of Twitter data comprises not only notification of the occurrence of an event itself, but also specific details conveyed about the event. This is not to say that simple awareness of events occurring is not useful, but rather that far more useful is specific information that informs and helps guide responses. Such information might include details of affected individuals and buildings, requests for specific or general help, or other time-sensitive information. Likewise as for the task of ED itself, there has been an increasing amount of research aiming at providing more fine grained classification of event related tweets beyond the simple relation of being relevant 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 is a critical problem in leveraging the potential power of such systems, stemming from a fundamental lack of fair and authoritative comparability between different techniques. This lack of comparability is the product of several interrelated factors. Foremost among these is the frequent unavailability of datasets used to conduct experiments. This stems variously from the use of bespoke datasets which are not made publicly available (see for example Liu et al. 2017; Tonon et al. 2017; Kanojia et al. 2016; Thapen et al. 2016; Hero 2016), among many more); to the removal of datasets owing to institutional and copyright issues (for example the Edinburgh FSD corpus of Petrovic et al. 2010); to various other 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 and filtering of otherwise available datasets (which renders it impossible to recapitulate the specific data utilised; e.g. Buntain et al. 2015) , and so on. Furthermore, there is often variation in terms of how the problem of ED and related information categorisation is defined, such that even two techniques using identical data cannot be compared like for like. Finally, owing to the often difficult to define quality of the outputs of such systems, there is even significant variation in the evaluation criteria and metrics used to assess systems (M. Weiler A. a. G. and Scholl 2015).

Thus, for parties interested in making use of such ED systems, there is no clear means of determining state of the art for their particular application. There is often in fact, no straightforward way at all of delineating between the efficacy of techniques, other than prominence in the literature, which can be misleading as to the quality and merit of the research itself. Furthermore, for research practitioners looking to advance capabilities in this area, there is no robust and shared baseline on which to build and improve. This renders advancing research in this area problematic, as there is no authoritative means by which to establish the superiority of an approach. Without such means to *authoritatively* assert advances to the field induced by new techniques, novel research has questionable value.

This paper therefore seeks to start the process of addressing this insufficiency. I select a technique that has been used for the task of both ED and related information types' classification, and attempt to replicate the reported results. The technique is chosen primarily for its demonstrated generalisability to the task of text processing (beyond just social media), and potential to extract multiple types of pertinent information. It is carefully re-implemented and experiments repeated as close as possible to those reported, on a publicly available dataset. Several modifications to the technique are subsequently explored, and finally the potential applications of the approach extended to novel crisis-related information types. It is hoped that this may then serve as a foundational baseline available to the community, against which other techniques can be accurately compared, and hence help clarify the path to improved capability in the area of ED.

The rest of this paper is structured as follows: first, an overview of the selected technique is provided. Secondly, details of the dataset used in the original paper and replication are explicated. Third, details of the replication process and experiments, including extensions to the technique, are reported. Finally, conclusions are made about the approach and problem area, informing suggestions for future work.

APPROACH

Selection of approach for replication was based on several factors. These included:

1. recent use in both research and applied context;
2. reputation of the technique for general efficacy at text processing tasks, indicating likely effectiveness to task of ED (and related sub-tasks);

3. availability of data used;
4. apparent leading/high performance based on reported results.

There is of course a risk that the technique does not represent the absolute best of extant approaches. Indeed this ambiguity forms a large part of the motivation for this research. Given the current lack of direct comparability, regardless of the technique ultimately selected, there is a dire need to establish a robust baseline, against which alternatives can be compared. Moreover, without said reliable baseline, there is no means of definitively determining this ranking, and hence this should be seen as a starting point upon which future research can build.

A plethora of different models have been used to perform ED[1]. Numerous time series based approaches have been tested, where events are detected as bursts in frequency of hashtags (Yilmaz and Hero 2016; Ozdikis, Senkul, et al. 2012; Costa et al. 2013), keywords (Hossny, Moschou, et al. 2018; Hossny and Mitchell 2019), and other derived features (Nützel and F. Zimmermann 2015; Comito et al. 2019). Similarly research has employed topic modelling (Zhu et al. 2017; B. Wang et al. 2017), various clustering based approaches (Alsaedi et al. 2017; Ozdikis, Karagoz, et al. 2017; De Boom et al. 2015), techniques leveraging the semantic web and curated knowledge bases (e.g. Tonon et al. 2017), and more recently neural network based approaches (Burel, Saif, and Alani 2017; Kruspe 2019), as well as various miscellaneous and approaches combining aspects of different strategies (Cordeiro 2012; Fang et al. 2016). Promise has been shown from all these families of techniques, but establishing the state of the art, or even exemplars of each sub-category, is difficult given the aforementioned lack of comparability interred in the research area generally (and which motivates this paper). Therefore, in selecting a technique to start the process of building a robust foundational baseline in ED, it was decided to select a technique which has not only shown promise within the area of ED, but one which has also shown general efficacy in other related text classification tasks, and hence is likely to provide strong performance to guide subsequent evaluation of, and comparison with, other techniques.

Burel et al. (Burel, Saif, Fernandez, et al. 2017) use a simple neural network architecture pioneered by Yoon Kim (Kim 2014)[2]. The model is designed as a generalisable sentence classification model, capable of being applied to a variety of downstream tasks. It has seen widespread use and success in a variety of tasks, and has formed the basis of many subsequent CNN based approaches to text classification (including for example 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).

The model takes as input word embeddings representing the source text. It comprises a single convolutional layer with 128 filters each of 3 different sizes (3, 4 and 5; i.e. 128*3=384 filters total) that extract features from the input embeddings. The convolution layers' outputs are then max-pooled, before being concatenated and passed through a softmax function to provide the final classifications. Burel et al. report parametrising the model to use dropout with probability 0.5 during training to mitigate over-fitting. They use ADAM gradient descent algorithm (Kingma and Ba 2014) and a batch size of 256. Training is reported as 400 iterations, though it is not mentioned whether early stopping is used to prevent over-fitting, nor whether optimal (validation) weights are restored. In recreating their approach, L2 regularisation was also tested, though it did not improve performance. Early stopping was also tested, and showed significant improvement as compared to continuing for the full 400 iterations reported. Burel et al. additionally use 5-fold cross-validation, though it is unclear what specific proportions of the different subsets of the corpus were used for train/validation/test sets (see data section for more details of the nuances around data folds).

Two variants of the architecture are tested in their paper, with different inputs to the neural networks:

1. using just word embeddings derived from the tweets' texts as inputs to the model.
2. using both word embeddings, as well as embeddings of "semantic concepts" contained within the tweets' texts as inputs.

For word embeddings, the widely used pre-trained GoogleNews word2vec (w2v) model [3] [4] is used, and for semantic concepts' annotation, they use the now defunct Alchemy API from IBM [5]. Fortunately (for the purposes of replicating the results reported in Burel) they make their semantic concepts' annotations publicly available (see data

[1]For a thorough overview of the variety of techniques levelled against the task I recommend any of the excellent surveys of Atefeh and Khreich 2015; Hasan, M. A. Orgun, et al. 2018; Imran, Castillo, Diaz, et al. 2018

[2]For which there are various implementations available, such as those listed at https://paperswithcode.com/paper/convolutional-neural-networks-for-sentence and https://www.kaggle.com/hamishdickson/cnn-for-sentence-classification-by-yoon-kim/data

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

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

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

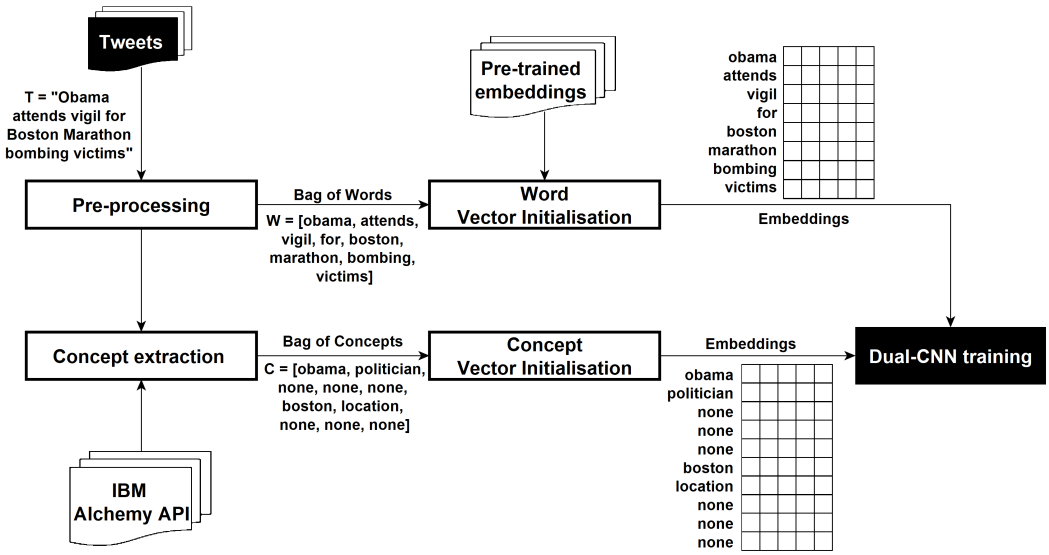

Figure 1. Overall pre-processing approach of Burel et al., combining both word and semantic concepts embeddings as input to a Dual-CNN model. (Modified from the original paper of Burel et al.)

Figure 2. 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']

section for more details) [6]. The obsolescence of the Alchemy API does mean though, that the specific concept annotation process cannot be recreated exactly for novel data (though a similar system is available within IBM Watson [7], and other alternatives providing similar functionality are available which could be used as substitutes [8]). Fortunately, as reported by Burel et al., and as confirmed in the replication results, the semantic concepts do not significantly improve results.

In both configurations (with and without semantic concept annotations), tweets are first pre-processed to clean and tokenise text. Unfortunately though, very few details of the specific text-cleaning or tokenisation process are provided. Tokens are then mapped to their respective pre-trained word embedding, creating for each tweet a matrix of word embeddings derived from the tweets' constituent tokens. Each matrix has dimensions 300 in one axis (corresponding to the length of the pre-trained embeddings' vectors), with the other dimension being equal in length to the highest number of tokens present in any of the tweets seen in the training set. Tweets with fewer tokens than this are padded to ensure consistent length. The process is illustrated by Burel using the diagram shown in figure 1.

In the semantic-concepts augmented model, a parallel operation extracts semantic concepts associated with named entities in the tweets, using the aforementioned IBM Alchemy API service. An embedding space is then initialised for these semantic concepts, and, as for the word embeddings, this results in a matrix per tweet of such concepts' embeddings. Since the majority of tweets do not contain mentions of any named entities, subsequently they are not annotated with "semantic concepts". Hence the semantic concepts embeddings' matrices are primarily uniform with embeddings representing "no semantic concept" of the source token. The semantic concepts' 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.

The semantic concept's embeddings are processed by an identical CNN as the architecture described above. The outputs of both the text and semantic concepts CNNs are concatenated immediately prior to being passed through the (final) softmax layer. Figure 2 from Burel et al. illustrates this process.

[6] Available at 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] Such as TextRazor, available at https://www.textrazor.com/

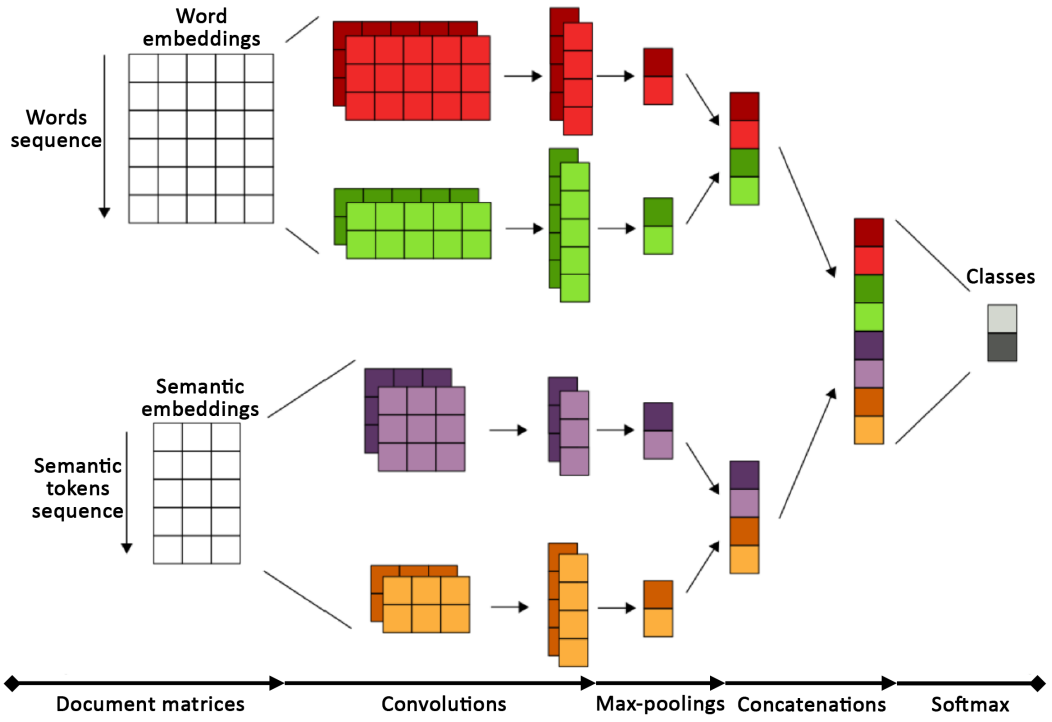

Figure 3. Visualisation of the dual-CNN method employed by Burel et al., showing separate ("dual") CNNs processing respectively the word & semantic concepts embeddings, before concatenation prior to final softmax layer. (Modified from the original paper of Burel et al.)

Burel et al. justify the choice of separate ("dual" in their wording) CNNs by virtue of the differing length of the tokenised texts produced in the 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 they used to perform semantic concept annotation, which introduces additional tokens denoting the class of the named entity *in addition to the named entity itself*, as seen in table 1. The single tokens "obama" and "boston" are annotated and introduce an additional token each, namely "politician" and "location". There is of course the possibility that modifying the semantic concepts' tokenisations to ensure length parity (by e.g. replacing named entities with just their class or token, whilst continuing to replace non-entity tokens with "none"), and hence enabling incorporation as a separate aligned channel would improve performance, but this has not been investigated here.

Finally, performance of the two architectures is assessed relative to 3 baseline approaches, respectively: Naïve Bayes (NB) classifier, Decision Trees, and Support Vector Machine (SVM). Unlike the CNNs however, the baseline methods are not given the same embeddings as input. Instead they are tested with TF-IDF representations of the tweets' texts. It's unclear why Burel et al. adopted this specific testing strategy, since it precludes the direct comparison of the baselines to their suggested approach by virtue of differing representations. Ideally the baselines should also be tested with the same inputs (both with and without the additional semantic concepts) to provide true comparison between the models' respective abilities to exploit those representations. Likewise as for aligning concepts' embeddings though, this has not been investigated here, but remains for future work.

DATA

One of the primary observable reasons for the lack of extant comparable research endeavours in ED is the dearth of high quality annotated data for the task. Many papers use bespoke datasets collected specifically for the purpose. Whilst perfectly valid for one-off and specific tests, the majority of these corpora are not made publicly available, and hence performance metrics reported in such papers become isolated by virtue of having no direct corollaries against which they may be compared. There may well be highly meritorious work conducted against such data, but it is of only limited worth to the community in general when it cannot be directly related to the work of others.

Despite this general approach and lack of data, there are *some* publicly available corpora with appropriate annotations to support this research. Fortunately, Burel et al. use one such corpus, namely the CrisisLexT26 (CLT26) [9] collection originally created by Olteanu, Castillo, et al. 2014 and extended by Olteanu, Vieweg, et al. 2015. Details of this dataset are provided.

CLT26 constitutes a collection of around 26000 tweets collected between 2012-13. These tweets are comprised of 26 different subsets, relating to specific individual crises, including both anthropic and natural instances. The collection is annotated with 3 different attributes, each with multiple categories:

1. `Informativeness`: 4-category label indicating whether a tweet is related to the event or not. Categories comprise "Related and informative", "Related but not informative", "Not related" and "Not applicable". This is the target used in ED classification tasks; 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" or "unrelated" to an event.

2. `Information Type`: 7-category label indicating the type of (crisis-relevant) 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 the types of information contained in event-related tweets. This is the target used by Burel et al. in their 3rd task of classifying the information type contained in tweets.

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

EXPERIMENTS

Burel et al. compare the performance of their three baseline models (Naïve Bayes, Support Vector Machine, Decision Tree) with their two CNN variants (with/without semantic concepts' annotation in a "dual" CNN) using CLT26 data. They test the models' ability to classify the tweets for three different targets, detailed in the "Replications" section below. By virtue of their intrigue and additional value that can be provided to crisis responders, I also tested several variants of their experiments, detailed in the subsequent "Extensions" section.

Replications

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

   Curiously Burel reports a final dataset for this target of 6703 instances, though despite contacting the author, no apparent reason for the discrepancy in count if forthcoming.

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

   Each Tweet belonging to an event specific subset is considered to be of that event type; e.g. all tweets included in each of the 3 subsets "2012 Colorado wildfires", "2013 Australian bushfire" and "2013 Brazil nightclub fire" are of the "fire" event type. This same approach is used in replicating this experiment, though it should be noted that this simple assignment process ignores the fact that many tweets in these event specific subsets have an "Informativeness" label of "Not related", and hence should perhaps more realistically be excluded.

   Finally, just as for the "Related/Unrelated" task, Burel et al. balances the dataset by undersampling. However, just as for the previous target, the figures reported by Burel et al. for the size of this balanced dataset are not consistent with what would be expected given their description. i.e. The smallest event type in

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

CLT26 is either Haze or Bombing, both of which comprise 1000 labelled tweets. Hence, undersampling the remaining 11 categories to the same size should result in a corpus of 12*1000 = 12000 tweets, not the 12997 reported by Burel et al.. Again, the authors have been contacted, but no explanation provided.

3. *Information Types*. 7-class label indicating the type of information contained in a tweet and included in the annotations provided by CLT26 as detailed previously. Likewise as for previous 2 targets, final dataset is balanced by undersampling, and, also likewise, the numbers reported in Burel do not match what one would expect when balancing the dataset in this way. The smallest class is "Not applicable" with 1138 tweets, and hence there should be 1138 * 7 = 7966 tweets, not the 9105 reported by Burel et al..

Extensions

1. *Embeddings variants*. Burel et al. use the publicly available pre-trained word2vec model trained on Google News articles. Since the linguistic style and constraints of news articles and tweets differ significantly, it should be straightforward to improve the performance of the model by using an alternative pre-trained model trained on twitter data. Several are available, including a recent release from Imran, Mitra, et al. 2016 which trains specifically on *crisis* tweets, and hence should be best matched to the CLT26 data. I test the three variants listed below:

   (a) *Godin Twitter 400d w2v* (Godin 2019): a 400d word2vec model trained on Twitter data, released by Frederic Godin[10].

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

   (c) *CrisisNLP 300d w2v*: a 300d word2vec model trained on crisis-related Twitter data, relreased by 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 the dataset through Twitter's API results in a significant drop in this total.

   Though not required for the approach of Burel et al., for any approach requiring access to the Tweets' metadata, re-hydration is essential, since the distributed form of CLT26 does not include much beyond the Tweets' texts and IDs, and hence this information must be obtained by re-downloading the Tweets from Twitter directly. Furthermore, dataset "rot" for Twitter corpora is common and it is often the case that subsequent use of Tweet corpora must work with a subset of the original. In order to gauge the impact of this reduction in corpus size, some experiments were repeated on the re-hydrated fraction of the corpus.

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

3. *Binary/mutliclass variants*. Testing the binary or multiclass equivalents of the targets' configuration tested in Burel, to assess relative complexity of each of the pertinent targets when moving from binary to multiclass setting and vice versa:

   (a) *Multiclass Informativeness*. Burel collapses the 4 class labels of CLT26 "Informativeness" annotation to binary "Related" / "Unrelated". I also test the full 4 class setting which would provide more nuanced information to crisis responders and journalists and permit more effective filtering in times of high tweet volume (to see i.e. just "Related and informative", rather than also including those that are "Related but uninformative", which may impede effective use of classifications).

   (b) *Binary event types*. Burel et al. test event types as a multiclass problem. I also test a binary variant which splits events into either natural or anthropic crises, indicating if the crises are natural disasters or the direct result of human actions.

[10] Available at https://github.com/FredericGodin/TwitterEmbeddings
[11] Available at https://nlp.stanford.edu/projects/glove/
[12] https://crisisnlp.qcri.org/
[13] Available at https://crisisnlp.qcri.org/lrec2016/lrec2016.html
[14] Rehydrated datasets, as well as all specific data configurations used for each experiment in this paper, are publicly available at https://github.com/j-m-crow/2020_crisis_baselines

(c) *Binary information types*. Binary version of the "Information types" annotations, with the intent of classifying the information types as either actionable (i.e. information that might be of benefit to crisis responders) or not (other information types which, though they may be related, do not aid or inform crisis response). Actionable categories comprise "Affected Individuals", "Infrastructure and utilities", "Caution and advice" and "Other useful information". Non-actionable categories comprise the remaining "Donations and volunteering", "Sympathy and support" and "Not applicable".[15]

4. *Information Source / Eyewitness*. The last extension to the work of Burel et al. is to include the additional annotation available with CLT26, namely "Information Source". The class is tested in both a multiclass (7 class) configuration and a binary variant, where all labels other than "Eyewitness" are collapsed to a single negative class. The motivation here is that most often it is eyewitnesses of crises that are able to provide the most up to date and accurate information on situational needs as they unfold. Hence, being able to accurately identify them could greatly aid in emergency response, and indeed there is increasing research in this area (see for example 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 of most interest from the both the replication perspective and in providing a wide baseline of the method and data to build upon in future research. Specific exclusions are noted in the various results tables.

IMPLEMENTATION

Numerous issues were encountered during implementation of experiments, concerning both the CLT26 corpus and the approach of Burel et al.. Hence, whilst attempting to replicate the method as closely as possible, numerous approximations and suppositions were necessary. Since these may impact the models performance, they are overviewed in precis below.

CrisisLexT26

- *Overlap between labelled crises*. As aforementioned, CLT26 comprises 26 subsets of tweets relating to individual crises. Owing to the overlap in both the keywords used to retrieve these tweets from Twitter API, and the time period (all events occurring in 2012-13), there are a number of tweets that appear in more than one subset. Furthermore, since the annotations of the tweets were conducted with respect to the collections for *individual crises*, the duplicate tweets appear with different annotations in the different sub-collections. For example tweet-ID '354439470801616898' appears in both "2013 Alberta Floods" subset, as well as "2013 Lac Megantic train crash" subset. The "Informativeness" and "Information Source" annotations are consistent between the two events, but not the "Information Type". For Alberta, it is labelled "Sympathy and Support", and for Lac Megantic it's "Other useful information".

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

- *Malformed semantic concepts annotations*. Just as for the original CLT26 corpus, the semantic concepts annotations released by Burel et al. also contain errors. A number of tweets contain no entry at all (note that this is distinct from not containing any annotated concepts - most tweets in the corpus contain no annotations for this, but are recorded in a specific format). More problematically, 3 entries contain annotations with a

[15]Note that there is no single notion of "actionability" adopted by the community that is applicable to all crisis types and crisis-responder roles (Ghosh et al. 2018). Rather, actionability is a product of numerous variables, including specific crisis type, the roles of responders making use of such information and application domain, as well as myriad other factors (Zade et al. 2018), and moreover is likely to evolve during the onset and development of crises (Munro 2011). The definition adopted here prioritises information as actionable in the immediate aftermath of crisis onset, and hence excludes "donations and volunteering" which, though also actionable, tend to be regarded as of lower urgency (C. Wang and Lillis 2020) than other information types which provide more useful and time critical information to responders in the crucial period immediately after the onset of a crisis event (Mccreadie et al. 2019).

large discrepancy between the number of tokens from the tweet text and number of tokens in the annotation. Tweet-ID 399804790227468289 for example has an annotation comprising 562 tokens (all "none") - more tokens in fact than there are characters permitted in a tweet [16].

Both tweets with mismatched annotation lengths and missing entries are clearly errors. Since the Alchemy API service no longer exists, it is also impossible to amend these errors. Hence, all tweets displaying such are dropped in these replications. Despite contacting to enquire, it is unclear whether these discrepancies were apparent during the experiments of Burel et al. or if they were introduced to the corpus later, prior to it's public release.

Burel et al.

- *Discrepancies in dataset sizes*. As briefly outlined above, numerous discrepancies exist in the dataset sizes reported in Burel et al.. Despite numerous attempts to infer possible causes for these discrepancies, and having contacted the authors, unfortunately no explanation has been provided.

- *Pre-processing*. Burel et al. provide only scant details of the pre-processing employed in their work. Hence, in recreating the cleaning and tokenisation of tweets, I have adopted a fairly standard approach that matches those few details. There remains the possibility though, that this significantly diverges from the original process. Specifically, I pre-process the tweets using the following procedure:

  1. data is cleaned (i.e. tweets with erroneous semantic concepts (as noted above) are removed, and any repeated tweets are de-duplicated).
  2. URLs are removed using simple regex pattern matching.
  3. any non-EN unicode characters are removed.
  4. all text is lowercased.
  5. tokens are extracted by using Keras' Tokenizer [17] using the default configuration.
  6. tokens are mapped to their pre-trained word embedding. Out-of-vocabulary tokens are handled as detailed below. Fine-tuning of embeddings during model training is allowed to improve model efficacy.
  7. semantic concepts (as provided by Burel et al.) are aligned with Tweets based on Tweet-ID.
  8. semantic concepts embeddings are initialised using Keras' Embedding layer. Likewise as for the token embeddings, these are fine-tuned during training to improve performance.

- *Cross validation*. Though Burel explicitly states the use of 5-fold cross validation, it is unclear how this is realised on the CLT26 data strata. It appears folds were applied to the aggregate dataset across all subsets, which could permit data from a single crisis to potentially appear in both the training and testing sets. Despite the inherent issues with this approach, since the aim is to validate results of Burel et al., this is the same strategy adopted in this research.

- *Model parametrisation*. Specific initialisation and parametrisation details of the CNN models and baseline approaches are not provided. Hence, significant parameter search was necessary to hone in on likely values providing optimal results. Limited by resources though, there is potential these could be improved further.

- *Out of vocabulary tokens*. Burel et al. do not explicate their strategy in dealing with out of vocabulary tokens that do not have entries in the pre-trained word2vec model used. Two common approaches are to use a fixed representation for all such OOV tokens, or alternatively to generate a random embedding for each new token according to the pre-existing embedding space distribution. Since the latter provides more granular token demarcation, this was used in the replication.

[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 1. Burel et al. reported results compared to replication results.

(Results in *italics* are those reported by Burel et al.. Those marked * equal or surpass Burel et al. (whilst those without achieve less than Burel et al.). For each metric (*P*recision, *R*ecall & *F1*) per-target top result is shown in **bold**, for both the full (unbalanced) and sample (balanced) dataset 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

RESULTS

Replications

Table 2 shows the results of the direct replication of Burel et al.'s experiments. Fortunately, the results are almost uniformly confirmatory of the results reported by Burel et al. and lend weight to the conclusions made therein. Minor exceptions exist where I was unable to match the results reported by Burel et al.. These are sufficiently small differences to be accounted for by simple lack of exhaustive parameter space search as dictated by pragmatic use of resources however, and not of concern. In several instances the results were surpassed (e.g. 0.012 increase in F1 for Information Type). Likewise as for the results I could not match, these differences are insignificant and likely the result in pragmatic trade-offs in training of both myself and Burel et al..

Concerning variance between the results, according to a two-tailed paired t-test measured over the 5 CV runs (of F1 measure), the differences between both CNN variants and SVM results are insignificant at $\alpha = 0.05$ and $\alpha = 0.005$, for both the Related/Unrelated and Information Types targets in either the balanced or unbalanced settings (independently). Conversely, both Naïve Bayes and Decision Tree results differences are shown to be significant (inferior) at both α values. Additionally, the difference between top performing models in each of the balanced and unbalanced settings were shown to be significant, again at both $\alpha = 0.05$ and $\alpha = 0.005$, highlighting the positive effect of class balancing this otherwise highly skewed dataset. Significance testing was not undertaken for the Event Types target owing to likely over-fitting of the model, as detailed in the Extensions results section below.

Extensions

Tables 3, 4, 5 & 6 show the results of the various extension experiments.

Embeddings variants

Table 3 shows the results of using different word embedding models in place of the Google News 300d Word2Vec model used by Burel et al.. Note that the Google News model results listed in this table are replication results, and not those originally reported in Burel. This is to ensure absolute parity and comparability between results in this table. Improvements are small but notable, with the Twitter-specific Godin w2v and GloVe embeddings providing the most consistent increase. Interestingly the Crisis-tweets trained model from CrisisNLP did not improve results as much as expected. Part of this is likely down to the pre-processing employed, particularly the removal of non-EN characters, which likely pre-disposed the tweets' contents to more *standard* linguistic style, and hence ensured the Google embeddings performed well despite domain mis-match.

Overall it appears that so long as the embedding model is sufficiently well (pre-)trained, the choice of specific model does not affect model performance too much. It would be interesting to test different approaches to pre-processing to see if these provided more distinct separation between the models' efficacies, as well as the potential to combine the

Table 2. Comparison of different embedding models for CNN and dual-CNN.

For each metric (*P*recision, *R*ecall & *F1*) per-target top result is shown in **bold**, for both the full (unbalanced) and sample (balanced) dataset configurations. Results marked * were not 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	*	*	*	*	*	*

multiple models as inputs, to leverage the combined information that each encodes, to see if there are more notable gains. Without further experimentation, it would seem that either the GloVe or Godin model is the best choice.[18]

Note that Event Types target has not been tested with further embeddings (in the rehydrated data configuration), since these results (being so high) suggest that the task in its current configuration is trivially easy for the models to solve. Indeed, Burel et al. note that the limited range of event types and specific event instances in the corpus devolves the task to one almost akin to keyword spotting, since for most crises in the corpus, identifying the type can be done simply by recognising the presence of one of a small set of keywords (as noted by Burel et al. for example, "77% of the tweets about meteorite falls contain the word meteor"). Hence, the model overfits to these specific keyword indicators, and would likely fail to generalise well beyond this limited context. It could be possible to ameliorate this issue by the removal of such indicators from the source texts and/or increasing the variety of specific event instances, but this has not been done in this paper owing to limited compute resource, and preference for testing the final annotation of CLT26 (namely Information Source) requiring pragmatic trade off.

Original vs re-hydrated data

Table 4 shows the results when using re-hydrated data rather than the original CLT26 corpus (compared to results in tables 2 & 3). As previously explicated, whilst not essential for the method employed by Burel et al. replicated herein, for numerous other ED (and related) techniques, the provision of Tweet metadata is **essential** for certain feature engineering. Hence, without including these re-hydrated data configured experiments, such techniques would remain incomparable to the results reported herein using only the original CLT26 data which omits such metadata. Therefore these results provide a reference against which to compare other research for which the re-hydration of CLT26 is *essential* (and which would therefore preclude comparison to the originally reported results). Tacitly they also provide guidance in the expected performance curtailment induced by the reduction in overall corpus size resulting from re-hydrating (as detailed previously).

Whilst certain metrics show increased performance (e.g. NB model achieves P=0.927 on re-hydrated data vs 0.898 originally), as expected there is an overall decline in performance. This is less extreme than expected, and is mostly limited to less than 5% difference. It should be noted though, that this form of testing does not provide sufficient

[18]Note that at the time the research was conducted, it was not feasible to incorporate contextual embeddings such as BERT (Devlin et al. 2019) and similar models into this extension comparing different word representations. Given the widespread demonstration of such embeddings' improvements in a variety of NLP tasks though, it is reasonable to expect that such embeddings would be at least competitive with the best performing embeddings here, if not superior. Whether such improvement would be a *significant* improvement though remains to be seen. In future work I plan to incorporate such contextual embeddings, as well as combinations of different sets of embeddings as noted above.

disambiguation of the models true predictive capacity. To truly assess the effect of the reduction in training corpus size, further experiments are required to test the trained models on entirely novel data. This would likely show a more marked reduction in performance owing to the far smaller variance in training tweets of the smaller corpus. These experiments are left for future work.

Table 3. Re-hydrated data results

For each metric (*P*recision, *R*ecall & *F1*) per-target top result is shown in **bold**, for both the full (unbalanced) and sample (balanced) dataset configurations. Results marked * were not 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	*	*	*	*	*	*

Binary / Multiclass variants

Table 5 shows the results of testing the two targets "Related/Unrelated" and "Information Type" in their multiclass/binary variant respectively. Event type, as mentioned above for re-hydrated data results, was not included in these experiments either owing to the lack of merit in doing so (the data being insufficient for reasonably representing the task).

Unsurprisingly, there is a marked dropoff in efficacy when moving from the binary related/unrelated configuration of Burel et al. to the more nuanced 4-class relatedness configuration afforded by CLT26 annotations. Whilst F1 peaks at 0.838 for the binary case (see 2), for the multiclass configuration this peaks at 0.664. Whilst this dropoff in performance is not so extreme as to render the model useless, the far lower F1 does indicate the tool would struggle to be effective in real world deployment. The performance could likely be improved by increased provision of higher quality and more diverse data (as discussed in conclusion). In the present scenario however, for those wishing to deploy live systems, I would recommend sacrificing the increased disambiguation of relatedness and instead opt for the more readily usable binary variant. There is additional potential that the performance of the binary variant could be made more useful by reconfiguring the positive class to include just the "Related and informative" class, and combining the "Related but not informative" class with the other two classes already treated as the negative class. This would ensure that any tweets actually classified as related do indeed contain informative (and therefore *actionable*) information, and further reduce overall noise in the deluge of social data surrounding high profile crises.

Just as unsurprisingly, the move from multiclass to binary configuration for information type shows an increase in performance (from peak F1=0.625 in the multiclass, to 0.85 in the binary). Corollary to the suggestion regarding deployment of binary configured relatedness classifier above, this marked improvement would suggest that crisis

Table 4. Binary / Multiclass variants results

For each metric (*P*recision, *R*ecall & *F1*) per-target top result is shown in **bold**, for both the full (unbalanced) and sample (balanced) dataset 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

responders would be better served by the simpler and higher performing binary information types. Being able to disambiguate actionable from non-actionable data with high precision being far more useful in time pressured responses than the far noisier and less informative multi-class information type classification alternative.

Information Source / Eyewitness

Table 6 shows the results using the "Information Source" annotation as classification target. Various experiments were run on this target, including the various embeddings variants, and using original vs re-hydrated version of the CLT26 corpus. This ensures that for all targets afforded by CLT26 (and which are more generally four commonly specified targets of useful information in crisis response information filtering) there exists herein parity of results provided. Furthermore, as previously detailed, there is increasing interest in being able to accurately identify the source of information, since it can provide such incisive means of filtering those (few) sources most likely to be able to provide the most timely and pertinent information as crises unfold. Particularly of interest in this regard is the binary configuration which poses the classification as either Eyewitness or not (i.e. any of the other classes). This is so since Eyewitnesses, by definition being co-located at the geographical location of an events occurrence, are best placed to provide accurate and up to date information.

The results are promising, and comparable to the performance of the model on the other targets in CLT26 (with the exception of Event Type which, as previously mentioned, is inherently easy to predict with this corpus and thus presents little usefulness). Choice of embedding model again seems to have relatively minor impact on overall performance. The binary configuration also shows far higher efficacy than the multiclass model. In fact, the binary configuration displays sufficiently high F1 to be of real usefulness, whilst the multiclass model (with peak F1=0.499) does not demonstrate efficacy sufficient to be of much real world use.

Interestingly, the peak performance is achieved by SVM using balanced (original, not re-hydrated) data. There are numerous possible explanations for this difference. Neural nets, requiring significantly greater amounts of data to train effectively, may simply be disadvantaged in this case by virtue of the limited size of the CLT26 corpus overall, and the additional dropoff in size resulting from balancing the class proportions. Additionally, the spatial aspect of CNNs may not be providing sufficient advantage over the SVM in the limited context of short tweets. Relatedly, the specific syntax of Eyewitness tweets may be such that the spatial characteristics of language that can be leveraged by CNN based approaches are simply not present. Further work is required to truly disambiguate this issue.

CONCLUSIONS & FUTURE WORK

This paper has demonstrated a number of interesting aspects of crisis-information classification on Twitter. Additionally it has highlighted key issues in the reproducibility of machine learning and social media focussed research which apply far more broadly. Foremost, the work has validated the baselines reported by Burel et al. through non-trivial replication of the work. Moreover, the various extensions to their work have further demonstrated the generalisability of the CNN design of Kim, and it's wide ranging efficacy in short-text classification tasks. These novel formulations (to both additional classification targets and test contexts) are of significant interest and practical

Table 5. Information Source results

For each metric (*P*recision, *R*ecall & *F1*) top result is shown in **bold**, for full (unbalanced) and sample (balanced) dataset configurations, per both original and re-hydrated data. Results marked * were not 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	*	*	*

use to both the research community, and practitioners looking to harness such systems, by providing more nuanced guidance for the appraisal and selection of such for real-world deployment.

The research also highlighted the need for more robust approaches to continued research in this area. There are numerous highly meritorious research avenues being pursued in this and related domains. However, without *authoritative* and *comparable* baselines against which to measure these novel research trajectories, there is no means by which we can measure the efficacy of this "progress". This paper provides one such baseline against which numerous other techniques can now be contrasted, in order that we can both better understand (indeed, in the first instance *establish*) the current state of the art, and then seek to improve thereupon.

Additionally, several areas of future work were identified. Most directly this comprises the extension of these results to more diverse extant approaches to the various problems. It also includes the potential to improve upon the highly effective, simple and generalisable CNN model variants tested herein, in both straightforward and more complex, nuanced ways.

In attaining the above contributions the work has also importantly highlighted discrepancies in the reporting of corpus statistics, and uncovered a number of errors in both modelling approach and corpus annotations. This has relevance far beyond ED, touching upon the broader challenge of reproducibility in machine learning research generally. Indeed, the primary difficulties in conducting the research stemmed from the lack of thorough and exhaustive reporting of process, and frequent obfuscation by omission in both data and modelling reporting, of crucial factors affecting the feasibility of independently recreating such.

The paramount importance of researchers in the community reporting the details and implications of their research as completely as possible cannot 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 in the creation or utilisation thereof, these must be reported in clear and cogent manner, whilst examining their potential impacts, in order that research validity is not undermined. Publicly available data sources must be utilised, either independently curated and made openly available, or through structured challenges such as the TREC incident streams track[19]. Where the potential for errors or ambiguities persists, these must be openly reported to not undermine subsequent research efforts, and enable their continued tackling and resolution by the wider community.

Significant time was expended in ameliorating these issues for this research, and I hope that demonstration of such detailed and thorough inspection required to do so, is invigorating to the community generally to focus more effort on such. Whilst it may not have the headline appeal of superficial advances on the "state of the art", ensuring the field continues to be supported by robust and verifiable methodologies and processes at its foundation is arguably of far more importance.

Finally, the research also helped to emphasise the need for more nuanced error analyses in this and similar work, and the dire need for higher quality, larger and more diverse training corpora. It is my intention to continue this research along both axes - increasing the quality and incisiveness of error analyses applied to crisis-classification, as well as seeking to improve the provision of data for this and related tasks. I hope that in so doing a collection of resources can be created and made available to the community, directly building upon the outputs already available from this research.

> "*Baselines are simultaneously one of the most valuable resources we have [...] They provide a sanity check against improvements, an easy avenue for the curious to begin to explore, and a potential foundation for future innovation to be built upon. Sadly, they're also one of the most neglected. [. . .] When we lose accurate baselines, we lose our ability to accurately measure our progress over time.*"
> - Steven Merity, 2017 [20]

[19] See http://dcs.gla.ac.uk/~richardm/TREC_IS/
[20] https://smerity.com/articles/2017/baselines_need_love.html

REFERENCES

Alcorn, S. (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: *2016 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 216–223.

Atefeh, F. and Khreich, W. (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: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. CIKM '17, 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 (SemDeep), at ESWC 2017*.

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

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

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

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

Comito, C., Forestiero, A., and Pizzuti, C. (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, Cham, pp. 1–41.

Costa, J., Silva, C., Antunes, M., and Ribeiro, B. (2013). "Defining Semantic Meta-hashtags for Twitter Classification". In: *11th International Conference on Adaptive and Natural Computing Algorithms (ICANNGA'2013)*, pp. 226–235.

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 (ICWSM)*.

De Boom, C., Van Canneyt, S., and Dhoedt, B. (2015). "Semantics-driven event clustering in twitter feeds". In: *Making Sense of Microposts* 1395, pp. 2–9.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *2019 Annual Conference of the North American Chapter of the Assocation for Computational Linguistics: Human Language Technologies(NAACL-HLT '19)*.

Diakopoulos, N., De Choudhury, M., and Naaman, M. (2012). "Finding and assessing social media information sources in the context of journalism". In: *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. Austin, Texas, USA: ACM Press, p. 2451.

Fan, M., Lin, W., Feng, Y., Sun, M., and Li, P. (2018). "A Globalization-Semantic Matching Neural Network for Paraphrase Identification". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM '18, pp. 2067–2075.

Fang, R., Nourbakhsh, A., LIU, X., Shah, S., and Li, Q. (2016). "Witness Identification in Twitter". In: *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. Austin, TX, USA: Association for Computational Linguistics, pp. 65–73.

Freitas, J. and Ji, H. (2016). "Identifying News from Tweets". In: *Proceedings of the 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. (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 (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 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. (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. (2016). "Multimodal Event Detection in Twitter Hashtag Networks". In: *Journal of Signal Processing Systems*.

Hossny, A. H. and Mitchell, L. (2019). "Event detection in Twitter: A keyword volume approach". In: *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*.

Hossny, A. H., Moschou, T., Osborne, G., Mitchell, L., and Lothian, N. (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 Proceedings of the Events and Stories in the News Workshop*, pp. 52–58.

Huang, Q. and Xiao, Y. (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. (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 The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pp. 507–511.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: Artificial Intelligence for Disaster Response". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14 Companion. New York, NY, USA: 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: *Proceedings of the 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: *Proceedings of the 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 (LREC)*.

Kalyanam, J., Quezada, M., Poblete, B., and Lanckriet, G. (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. (2016). "Civique: Using Social Media to Detect Urban Emergencies". In: *Very Large Databases (VLDB) 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: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751.

Kingma, D. P. and Ba, J. (2014). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR '14)*.

Krumm, J. and Horvitz, E. (2015). "Eyewitness: Identifying Local Events via Space-Time Signals in Twitter Feeds". In: *23rd SIGSPATIAL International Conference (GIS '15)*.

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

Liu, X., Nourbakhsh, A., Li, Q., Shah, S., Martin, R., and Duprey, J. (2017). "Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data". In: *2017 IEEE International Conference on Big Data (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 Conference on Information Systems for Crisis Response and Management (ISCRAM 2019)*, pp. 691–705.

Morstatter, F., Lubold, N., Pon-Barry, H., Pfeffer, J., and Liu, H. (2014). "Finding Eyewitness Tweets During Crises". In: *ACL 2014 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: *Proceedings of the 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. (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: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*, pp. 994–1009.

Ozdikis, O., Karagoz, P., and Oğuztüzün, H. (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 (WOSS 2012)*.

Pagolu, V. S., Challa, K. N. R., Panda, G., and Majhi, B. (2016). "Sentiment analysis of Twitter data for predicting stock market movements". In: *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 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: *2016 15th Annual International Conference on Security and Management (SAM'16)*.

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: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by S. Kraus. ijcai.org, pp. 3238–3245.

Pennington, J., Socher, R., and Manning, C. (2014). "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.

Petrovic, S., Osborne, M., and Lavrenko, V. (2010). "Streaming first story detection with application to Twitter". In: *Human Language Technologies - The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*.

Repp, Ø. and Ramampiaro, H. (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: *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–6.

Sen, A., Rudra, K., and Ghosh, S. (2015). "Extracting situational awareness from microblogs during disaster events". In: *2015 7th International Conference on Communication Systems and Networks (COMSNETS)*, pp. 1–6.

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: *The Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-18)*.

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: *2019 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 Conference on Information Systems for Crisis Response and Management ISCRAM 2012*.

Tanev, H., Zavarella, V., and Steinberger, J. (2017). "Monitoring disaster impact: detecting micro-events and eyewitness reports in mainstream and social media". en. In: *14th International Conference on Information Systems for Crisis Response and Management ISCRAM 2017*, p. 11.

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 2017 (ESWC '17)* 10250 LNCS, pp. 138–153.

Truelove, M., Khoshelham, K., McLean, S., Winter, S., and Vasardani, M. (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. (2017). "Testing the event witnessing status of micro-bloggers from evidence in their micro-blogs". en. In: *PLOS ONE* 12.12. Ed. by E. Ito, e0189378.

Tsapeli, F., Bezirgiannidis, N., Tino, P., and Musolesi, M. (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: *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 515–522.

Vargas-Calderón, V., Parra-A., N., Camargo, J. E., and Vinck-Posada, H. (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. (2017). "A Hierarchical Topic Modelling Approach for Tweet Clustering". In: *9th International Conference on Social Informatics (SocInfo 2017)*, pp. 378–390.

Wang, C. and Lillis, D. (2020). "Classification for Crisis-Related Tweets Leveraging Word Embeddings and Data Augmentation". In: *Proceedings of the Twenty-Eighth Text Retrieval Conference (TREC 2019)*.

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: *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2384–2391.

Weiler, A., Grossniklaus, M., and Scholl, M. H. (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*. Ed. by S. Maneth. Cham: Springer International Publishing, pp. 108–119.

Yilmaz, Y. and Hero, A. (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: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW, pp. 1–18.

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 Conference on Information Systems for Crisis Response and Management ISCRAM 2018*.

Zhang, H., Ma, F., Li, Y., Zhang, C., Wang, T., Wang, Y., Gao, J., and Su, L. (2018). "Leveraging the Power of Informative Users for Local Event Detection". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 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, Berlin, Heidelberg, 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.