presented at 7th IFIP WG 5.15 International Conference, ITDRR 2022, Kristiansand, Norway, October 12–14, 2022

Learning early detection of emergencies from word usage patterns on social media

Carlo A. Bono[1][0000−0002−5734−1274]
Mehmet Oğuz Mülâyim[2][0000−0002−3993−5597]
Barbara Pernici[1][0000−0002−2034−9774]

[1] Politecnico di Milano, DEIB
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
{carlo.bono,barbara.pernici}polimi.it
[2] Artificial Intelligence Research Institute (IIIA), CSIC
Campus UAB, 08193, Cerdanyola del Vallès, Spain
oguz@iiia.csic.es

Abstract. In the early stages of an emergency, information extracted from social media can support crisis response with evidence-based content. In order to capture this evidence, the events of interest must be first promptly detected. An automated detection system is able to activate other tasks, such as preemptive data processing for extracting event-related information. In this paper, we extend the human-in-the-loop approach in our previous work, TriggerCit, with a machine-learning-based event detection system trained on word count time series and coupled with an automated lexicon building algorithm. We design this framework in a language-agnostic fashion. In this way, the system can be deployed to any language without substantial effort. We evaluate the capacity of the proposed work against authoritative flood data for Nepal recorded over two years.

Keywords: Social Media · Disaster Management · Early Alerting.

1 Introduction

The use of social media as a data source during emergencies has been largely investigated in the last decade [6]. Social media are often used to document ongoing events, and can provide real-time information in the form of text and media. As a consequence, social media platforms can deliver situational awareness and support an effective response to large-scale disaster events, helping to mitigate losses. In order to be valuable, data extracted from social media must be as timely and accurate as possible [20]. In our previous work [3], we focused on the automatic derivation of qualified and geolocated evidence to be delivered to responding organizations. In this work, to complement and expand previous results, we specifically focus on the automatic detection of events. A prompt event detection mechanism can be used to trigger or preempt tasks related to emergencies.

The guiding idea of the present work is to build a tool that, with minimal supervision, can detect the onset of emergency events in near-real-time. The input data for such tool are word mentions on social media. This objective is achieved through a data-centered approach, aimed at predicting if an emergency event is ongoing or imminent. A set of keywords linked to the onsets of a certain class of events is retrieved. Then, a predictor making use of the mentions of these keywords over time is built. Both steps are performed automatically and offline, prior to the events. Together with event identification, we also evaluate how to derive an indication of its magnitude, by estimating the number of recent incidents. This approach is meant to be timely, lightweight and general purpose. Multilingual support is achieved by designing a language-based approach, whose construction is both automated and language-independent, so that it can replicated with ease on any language. This approach, as a byproduct, also enhances the recall of the system, since it produces a language-specific lexicon tailored to a class of events. This characteristic is also critical for obtaining representative results in subsequently activated stages, when social media posts are crawled and inspected searching for informative evidence.

The paper is structured as follows. Related work is framed in Section 2. The data and methods utilized are documented in Section 3, while Section 4 presents the experimental results. Discussion is proposed in Section 5, and future work is outlined in Section 6.

2 Related work

Emergency event detection can be performed using direct sensor data, remote sensor data or indirect data. Sensor-based approaches are possible for a number of emergency events. In the case of floods, water level data and rainfall intensity are usually measured. While direct-measure approaches can be accurate, their deployment is rather costly, and poses geographical coverage issues. For example, survey data suggests that the majority of river basins are equipped with insufficient gauging stations for observing water level, streamflow and rainfall [16].

Remote observation conducted through satellite imaging and active/passive sensing has been widely adopted for flood monitoring and tracking. Approaches based on remote sensing often exploit machine learning capabilities for task automation [12]. For example, systems like FloodAI [13] utilize Synthetic Aperture Radar (SAR) imagery and machine learning to perform remote flood analysis. Among the main limitations of this type of approaches there is their computational complexity, which therefore requires delimiting the area of interest before their application, and therefore require some other early detection mechanism in order to be activated.

The use of lexicons in social media emergency management has been investigated in [15]. Multilingual lexicons usually aggregate keywords of interest from a number of supported languages, for example 60 in [5] and 32 in [14].

The use of indirect signals extracted from social media, in particular Twitter, for event detection have been discussed in many scenarios. For example, early warning systems for earthquake events have been extensively studied [17, 2]. An interesting solution for flood event detection using exclusively social media data is provided in [4]. The authors leverage the arrival time of flood-related tweets using a fixed dictionary of terms for social media crawling. The viability of a global flood monitoring system with self-activation capabilities has been explored in [11], suggesting recall issues when the approach is purely based on social media data. Additionally, the advantages of combining sensor and social media data for early warning purposes are analyzed in [18], while the potential of integrating remote sensing and social media specifically for early flood detection is explored in [8].

In the present work, we combine a self-building, dictionary-based approach with a machine learning setup, and apply it to flood event detection. Our methodology is agnostic to both the specific kind of event and the language of choice, making it suitable for a multi-language approach and adaptable to other emergency events. To the best of our knowledge, such end-to-end approach to event detection is novel. This study complements previous work described in [3], implementing part of the future work envisioned therein towards learning search keywords dictionaries and leveraging machine learning for event detection. Neural network approaches for the supervised learning setups proposed in this paper draw inspiration from the review work in [7] and in particular the fully convolutional network proposed in [21].

3 Data and methodology

The driving goal for the current work is building a language-centered system that can detect events using social media data for early alerting purposes. From a broad perspective, two main ingredients are needed to build such a system with a data-driven approach: signal data and a ground truth, which together constitute the training data for our system. Signal data is, in this case, a representation of reality coming from social media. To this extent, we use a minimal representation: the count of the usage of selected words over time. As ground truth, we can use historical validation data on past events, as described in Section 3.1. Once a proper training set is assembled, supervised learning approaches for the automated detection of event onsets can be assessed. We give a description of the approaches we have evaluated in Section 3.3.

Regarding the signal data, consisting of word counts, the main focus is to derive automatically a dictionary of words that are significant to events of interest. If the system has to support multiple languages, and the queries to social media are to be posed as textual queries[3], a sensible solution is to use language-centered word dictionaries. For the sake of automation, the burden of building these dictionaries should also be delegated to machines as much as possible. The

[3] Searching on social media is usually done with keywords, in combination with logical and advanced operators.

automatic construction of event-specific queries counterbalances the combined lack of language and domain-specific knowledge, which is almost inevitable in a general purpose system. We describe our approach for dictionary derivation in Section 3.2.

The overall approach is illustrated in Fig. 1. Starting from a small set of terms, the initial dictionary is expanded using both offline and online methods informed by the ground truth. Once the dictionary is set, the time series corresponding to each keyword count are fed to a supervised learning algorithm, together with the ground truth. The resulting model is able to perform predictions on unseen data and detect an ongoing event.

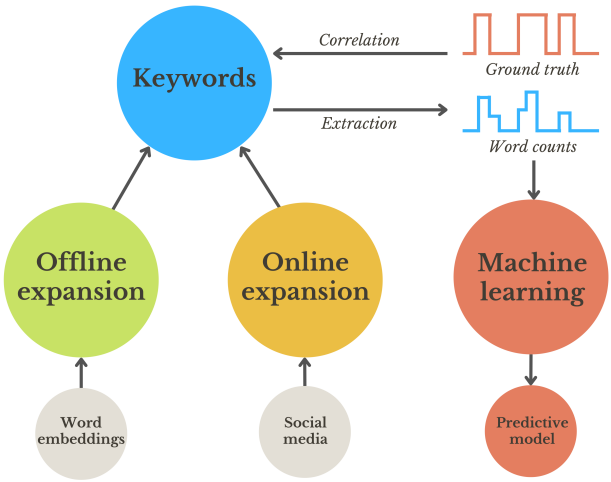

Fig. 1: Overall schematic of our approach to dictionary construction and model fitting.

3.1 Data and ground truth

We use the counts of keywords over time, extracted from a social media platform, as a signal describing the reality. The experiments performed are based on Twitter data. This approach is grounded on the assumption that some words are remarkably used when a particular event happens. Words that show this behaviour are said to be "correlated", according to some quantitative measure, with the event type of interest. Single keywords are a simple feature compared to the complexity of natural languages, yet they are at the core of the queries that can be posed to social media and, for practical purposes, keyword dictionaries are frequently exploited in emergency management applications on social media.

Utilizing the word count brings about some advantages. First of all, it is a conveniently compact feature consisting in a list of integers. This feature can be easily computed. Twitter API also exposes a `v2 Tweet counts` endpoint, which

returns the requested counts with configurable granularity, without returning actual posts. This option makes data interchange and processing negligible. Full archive search required for training is available with the Academic Research access; on the other hand, access to recent data –which is suitable for real-time operation– is generally available. Data can be returned at different time granularities depending on application needs. We mainly experimented with counts aggregated at hourly level and windows of 7 days, which coincide with Twitter's "recent search" visibility. An example depiction of the resulting word counts is shown in Fig. 2.

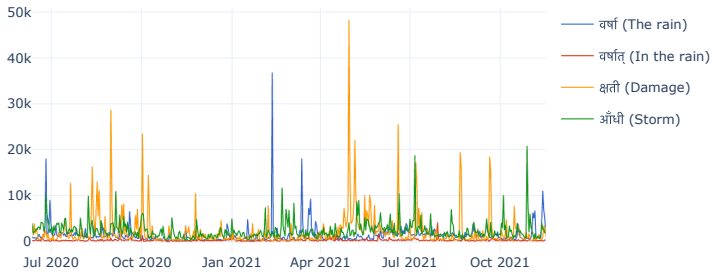

Fig. 2: Example of word count time series from Twitter.

When some event occurs, keywords related to the ongoing event are not always related with other events of the same kind. An obvious example is mentions of locations. In this work, we focus on the derivation of general purpose dictionaries, without investigating approaches aimed at dictionary adaptation to a single, currently ongoing event, such as the one proposed in [1].

Regarding ground truth data, we mainly experimented with data extracted from the *Global Disaster Alert and Coordination System*[4] (GDACS) [19], which is designed to alert the international community during sudden-onset disasters. Since our case study is based on Nepal, we focused on the 12 events reported over the last two years in the country.[5] As the investigation developed, we also considered the data sources reported in [10]. Some of the listed datasets were not applicable to this study. The comparison with available sources highlighted a data quality and definition issue that is discussed in Section 4 and can be observed in Fig. 10. Moreover, we also used incident reports from the Nepal Disaster Risk Reduction Portal[6], both for comparison and training purposes.

[4] https://www.gdacs.org
[5] We did not use previous reports since, ostensibly, the data collection process changed at some point.
[6] http://drrportal.gov.np/

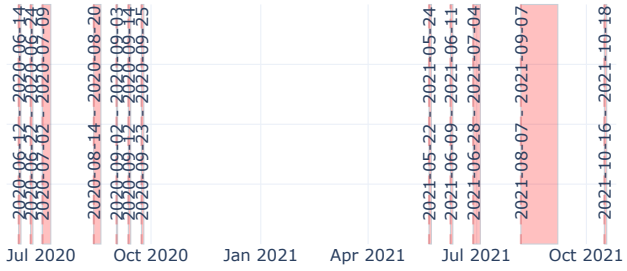

Fig. 3: GDACS time ranges for flood events in Nepal.

3.2 Automatic dictionary building

In the given context, a dictionary building method should have some desirable characteristics. It should be as automatic as possible, in order to be replicable on different languages. It should also rely on widely available resources. Finally, it should leverage accessible ground truth to assess the usefulness of the candidate keywords. In this section we detail how to achieve such characteristics. The final goal is to obtain a representative set of keywords that can guarantee a high recall, with reasonable specificity, when used to query a social media platform.

Dictionary expansion An initial dictionary is created with a small number of "seed" keywords. These keywords are generically related to the event of interest (e.g., "flood" in the desired language). Starting from these few keywords, the first step is to expand the dictionary with candidate keywords that are related to the initial ones. Such relatedness can originate from different sources, such as language models (e.g., bigram models) that are usually available for most languages, or even search engine data such as Google Trends data[7]. We choose to use non-contextual, language-specific word embeddings for the expansion, since they roughly capture semantic proximity and they are generally available for most languages. For each seed keyword in the initial set, we add the top N most similar words to the set, and then perform the same expansion a second time. A visual representation of the expansion process is given in Fig. 4.

We also want to filter out obvious outliers. To this end, a light manual filtering is performed on the candidate keywords, using automatic translation and an interactive dashboard.

Correlation with event onsets Given a set of keywords, we want to retain the most significant candidates. To compute a measure of significance, we use Pearson and Kendall coefficients. We compute these coefficients between time-lagged shifts of each word count vector and the time series of the event onsets. We take care of shifting the time series in one direction only, in order to avoid measuring

[7] https://trends.google.com/

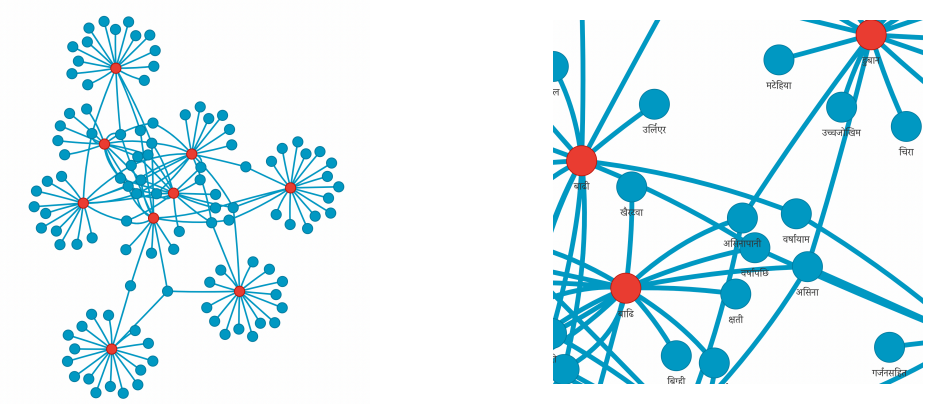

Fig. 4: Offline expansion of the seed dictionary using Nepali word embeddings.

inverse causality between events and words. For each word, we average the maximum coefficients. We then retain only positively correlated candidates. Measures of correlation calculated in this way are usually low. There are a number of reasons for this. Some are structural, in the sense that we are looking for keywords correlated with the events onset, while they could instead be correlated with the overall unfolding of the events. Some others are due to the nature of the events, which is variable and unpredictable. An emblematic example is provided in Fig. 5. The keyword "landslide" shows distinctive spikes in correspondence with some of the onsets, while in some others the spike is delayed or missing altogether.

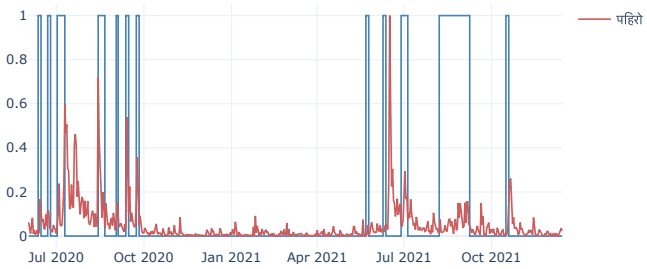

Fig. 5: Normalized hourly count for "landslide" and GDACS flood events.

Data sampling Selected keywords, seemingly correlated with the events onsets, are then used to build a query consisting in the logical OR ($\vee$) of all the keywords. We use this query to download matching posts in correspondence with the onset days, namely *positive* days. In this way, unseen keywords originating from online data now become accessible. However, it would be impractical to get all the historical counts for all the words to compute the correlation as described. A

different strategy is then used. We also sample tweets from days that are far from the events by a given interval before and after. We call these days *negative* days. This second query is composed by the OR of keywords randomly sampled from common terms in the language of interest.[8] The time masking used for positive and negative sampling is illustrated in Fig. 6. All downloaded tweets are then projected to a vector space defined by the keywords witnessed at least K times in the samples coming from the positive days. Then, the χ^2 statistic is used to rank the most significant terms. The top ones are added to the dictionary if they also achieve a significant correlation with the event onsets.

Fig. 6: Event onsets from GDACS, and time windows far from the events.

A validation dashboard, visible in Fig. 7, is finally used to filter out spurious candidates, such as location names, which are event-dependent.

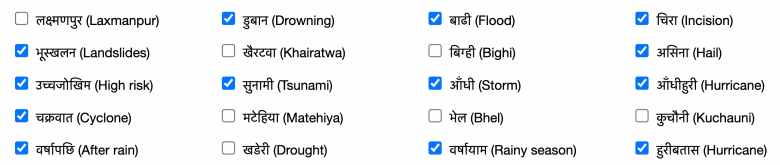

Fig. 7: Manual dashboard for keyword filtering.

The whole procedure can be optionally reiterated, since the OR query changes when the dictionary contents change, refining the retrieved terms up to convergence.

3.3 Supervised learning

Once the dictionary is set, the time series for the selected keywords are used as input data for a supervised classification problem. We first evaluated a classification setup, using GDACS data as a ground truth. We then studied a regression setup using incident data related to our case study.

Features and preprocessing For each keyword in the dictionary, we obtained a dense vector of counts at a requested time granularity. We mostly experimented with hourly data. Some machine learning algorithms work better with

[8] This is mandatory on Twitter since a query consisting of only stopwords is rejected.

a proper scaling of the input features. Also, different data preprocessing choices could influence the final performance. Different classical feature preprocessing are evaluated, such as min-max scaling, Z-score normalization and median-IQR scaling. For each experiment, the selected scaling is applied individually to each word's time series. Additionally, quantile-based signal quantization is applied to the scaled time series. The choice of the preprocessing is deferred to the experimental stage.

Classification setup The first investigated setup is a classification task. In this setup, each vector is related to a specific time index (e.g., an hour of a specific day) and contains normalized and quantized word count data over the past week. First, we create a vector for each index within the time window of three days before and two days after an onset (both dates inclusive). These vectors form the *positive* cases. We expect positive cases to encompass the data of anticipated events (e.g., rains) and early reactions. We do not take into account the remaining dates during the events (i.e., third day after the onset and onwards) since we are interested in alerting. Then, we create vectors belonging to time windows that are "far from the events", as seen in Fig. 6, and these vectors form the *negative* cases.

Vectors indexed by the same time are grouped, so that the resulting input data is a matrix obtained concatenating count vectors of equal length, corresponding to a multivariate time series. Since the relations between the magnitudes of the time series are lost during preprocessing, a vector containing mean, standard deviation, maximum and minimum is computed over the same time range on the raw counts. For each time index, the inputs for the classifier are a $n \times t$ matrix plus a supplementary $n \times 4$ vector, where n is the dictionary cardinality and t is the size of the time window.

We assign class weights inversely proportional to the number of negative and positive instances. Evaluation of the models is based on standard classification metrics, namely, precision, recall, F_1 score and AUC.

The base model for the experiments was the Fully Convolutional Network (FCN) approach proposed by [21]. An exploration of the effect of varying the number of layers, units, kernel size was conducted.[9] Values for learning rate (LR), LR reduction, batch size and early stopping were calibrated during the experiments. Dropout layers were added at the first layer and after the global average pooling layer. Also, the ResNet architecture proposed by the same authors has been tested. Experiments are performed in a leave-one-event out fashion, repeating each experiment 10 times in order to average stochastic effects. Adam [9] is used for gradient descent optimization, and categorical cross-entropy as a loss function. Validation accuracy is used as a metric for early stopping. The validation loss was computed on a random 33% of the training cases, for each fold.

Regression setup Following experimental observations on the classification results, we applied an analogous deep neural network setup to a regression task to

[9] By mixing manual exploration and automatic exploration using KerasTuner.

estimate the occurrence and impact of flood incidents. Incident count data for the region of interest are aggregated by day, regardless of the specific location within that region. A multi-headed network setup for predicting life and monetary losses has also been evaluated. In the regression setup, the full data is fed to the network, without masking days. This is motivated by the nature of the ground truth data. As for the classification case, possible anticipation and delay effects between word use and actual events have to be accounted for. Therefore, we chose to replace each data label by a 72-hour arithmetic average, centered in each point, similarly to the classification setup. Mean squared error (MSE) is used as evaluation metric.

Experiments are performed in a 5-fold cross validation fashion, where each fold contains data points that are contiguous in time. Again, each experiment is performed 10 times and the results are averaged. Hyperparameter choices are deferred to experimentation. MSE is used as loss function and evaluation metric. The validation loss is computed on the last 50% of the training cases, for each fold.

4 Experimental results

We first analyzed the classification setup choosing flood events in Nepal as a case study. This is done in continuity with our previous work, in which a flood event in Nepal was analyzed [3]. We had selected this event since the United Nations Satellite Centre (UNOSAT)[10] was activated to support it. We generalized the validation of our models to all recent flood events in Nepal. Moreover, since Twitter penetration rate is low in Nepal, it also poses a challenging test. Based on the results, we then analyzed the regression setup using the same input data and flood-related incident data reported by the Government of Nepal as ground truth. In both cases, roughly the last two years of data were analyzed.

4.1 Data gathering

We applied the procedure described in Section 3, starting from two flood-related Nepali words (बाढी and बाढि). Publicly available word embeddings for Nepali[11] were used, adding $N = 10$ similar terms per word and performing the expansion two times. After term expansion, manual filtering, and tweet data downloading[12] in order to get real-world keywords, we conducted the correlation analysis described in Subection 3.2 which led to 41 Nepali words that where related to the flood event onsets extracted from GDACS. For each data point, indexed by each round hour in the time series, a one-week count time series was extracted for each word, leading to a $16632 \times 41 \times 168$ input matrix, with each data point consisting in a multivariate time series represented by a 41×168 matrix. The

[10] https://unitar.org/sustainable-development-goals/united-nations-satellite-centre-UNOSAT

[11] https://github.com/rabindralamsal/Word2Vec-Embeddings-for-Nepali-Language

[12] Approximately 5 million tweets sampled from positive and negative days.

supplementary 41×4 vector containing descriptive statistics was computed for each time window and added to the corresponding multivariate time series.

4.2 Preprocessing and hyperparameter selection

To select the most suitable configuration for feature preprocessing, we repeated the classification experiments averaging the output probabilities over 10 runs for each configuration. Output probabilities correspond to the softmax probabilities of being a positive case. In Tab. 1 the Brier score for the average outputs is reported. Based on this analysis, we chose to use MinMax preprocessing and 200 quantization levels.

Normalisation type	Brier avg	Brier stdev
None	0.114	0.003
MinMax	0.069	0.003
MinMax (qcut 100)	0.087	0.002
MinMax (qcut 200)	0.068	0.003
MinMax (qcut 300)	0.070	0.003
Standard	0.083	0.002
Standard (qcut 100)	0.098	0.006
Standard (qcut 200)	0.100	0.005
Standard (qcut 300)	0.083	0.003

Table 1: Preprocessing selection with Brier score, average scores over 10 runs.

Hyperparameter tuning for the classification setup was done through KerasTuner and manual testing, focusing on networks with few layers and units since bigger network configurations showed overfitting behaviour. In Fig. 8, an exploration of the number of units per layer and kernel size is reported, focusing on the regression setup. While such exploration is limited by the fact that only a parameter at a time is studied, the analysis shows that the loss is not strongly influenced by the parameter itself, up to some range. This is confirmed by the fact that further random KerasTuner exploration did not lead to models that achieve better performances.

4.3 Results with classification

Relatively small networks show a reasonable performance in the classification task. For reference, the output probabilities for a 2-layer, 32 units network are reported in Fig. 9. Using 0.2 as a probability threshold for the output layer, this network achieves a 94% precision and a 74% recall. Magnitude-related statistics did not prove to be useful and were discarded from input.

Recall is computed over 1,728 positive cases, evenly distributed over 12 events. Focusing on the ability to intercept the events, such model looks to be able to get all the available event onsets, at the cost of some false positives.

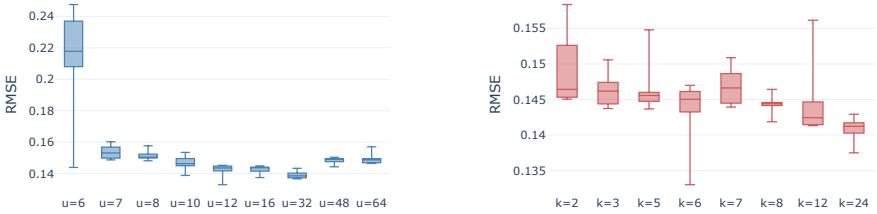

Fig. 8: Response to number of units (u) per layer and kernel size (k) for 1D convolution.

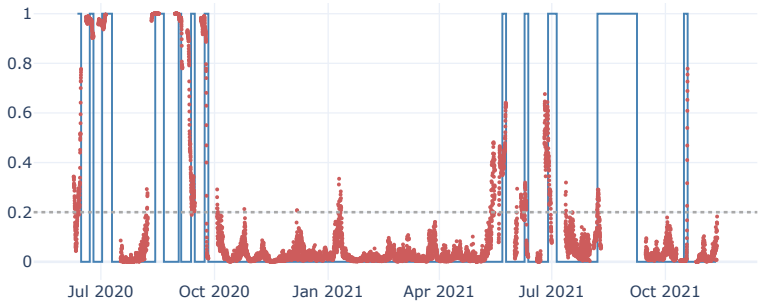

Fig. 9: Probability outputs (red) for event onsets of a 2-layer, 32 units CNN. Blue lines show event days. Dashed line marks the threshold for positive prediction.

4.4 Reference issue

However, the output probability is not evenly distributed over the output events. Fig. 9 highlights a different prediction behaviour between older and newer events. After evaluating an extensive number of network configurations with no significant enhancement, we conducted a comparative analysis of available ground truths for flood events in Nepal, using accessible datasets listed in [10].[13] In particular, we compared GDACS data with EM-DAT[14] and the Global Active Archive of Large Flood Events[15]. As shown in Fig. 10, the event onsets reported by different data sources are inconsistent.

In an effort to understand the relation between these ground truths, we also compared event data with incident data coming from the Nepal Disaster Risk Reduction Portal (NDRRP). As it can be appreciated in Fig. 11, when the classifier is properly fitting the data it is able to predict days with actual incidents as positive cases, as opposed to days marked as onsets.

[13] We did not use Global Flood Monitor data described in [4] since it does not contain certified data, and we were not able to obtain NatCatSERVICE data from Munich Re.

[14] The International Disaster database, https://www.emdat.be/

[15] https://floodobservatory.colorado.edu/Archives/index.html

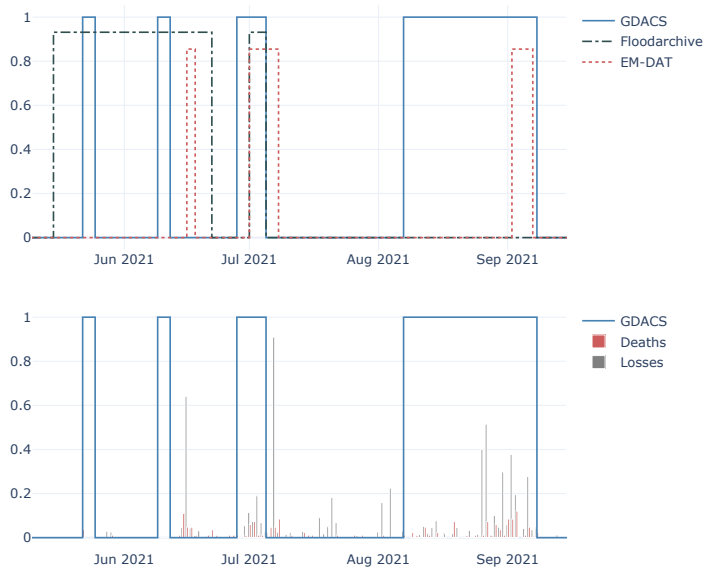

Fig. 10: Comparison between GDACS, EM-DAT, Floodarchive and NDRRP event data.

4.5 Results with regression

The previous observations on classification results suggested that incident data could be a reliable target variable. Since incident data is not a binary value, the learning setup was changed to a regression. In Fig. 12 the prediction of the normalized number of events is reported. The predictions are aggregated from a 5-fold cross-validation setup and are made on unseen data. The corresponding MSE is 0.141.

The experiments with the ResNet network, configured with a comparable number of units and kernel size, did not enhance the results. Moreover, the ResNet approach was about 20 times slower to train in our experiments. Experiments with supplementary network heads for learning life and monetary losses as additional dependent variables –together with the incident count– did not lead to better performances with respect to the one-headed network.

5 Discussion and limitations

Both the classification and regression setup showed usable results in terms of generalization capabilities and overall performance metrics. Since the approach is very compact and the computation negligible, the system can be considered ready for an experimental deployment. No particular functional or non-functional constraints are given, apart from the input vectors being related to the last week of data.

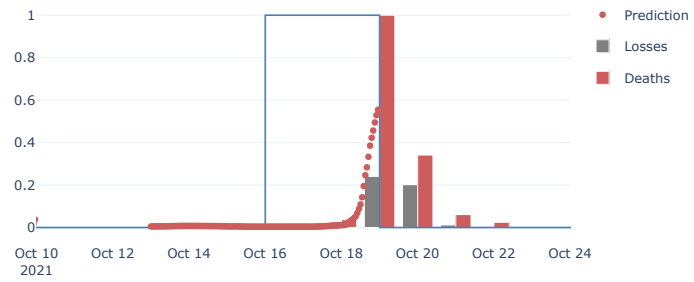

Fig. 11: Classifier generalization on the last flood event reported by GDACS.

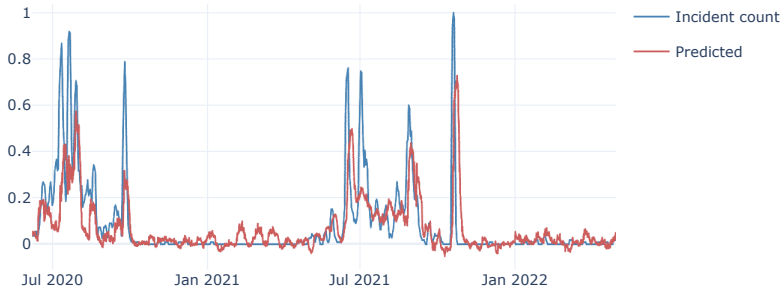

Fig. 12: 5-fold cross-validation results for incident count regression.

While the performances of the classifier setup are fit for event detection, the observations presented about the ground truth make the use of the resulting model less promising. Specifically, even if the events themselves are correctly identified, we observe variation in the probability output of the network depending on the event, partly due to the nature of the ground truth itself. Notwithstanding this limitation, the generalization capability of the model appear to be good.

In the regression setup, under certain circumstances the peaks of the predicted values show a time lag with respect to the reference values. Since the effect only happens for some events, it is more likely to be a real-world tailing effect rather than a byproduct of the chosen preprocessing. This does not invalidate the functionality of the approach since, even if the predicted peaks are possibly shifted, the slope and the absolute value of the prediction are still satisfactory as a trigger. Given the nature of the results on the negative regions (Fig. 9 and 12), a postprocessing technique to mitigate false positives could be necessary, especially in highly time-sensitive applications. It is also worth noticing that, in the context of emergency alerting, the cost of false positives is usually negligible compared to false negatives.

Moreover, since the dictionary-building phase has been consistently tested, we believe that applying this approach to other languages is practicable with no particular effort. However, regarding the regression setup, a suitable ground truth should be obtained for the regions of interest. While the proposed models

appear to be able to generalize well, the availability of a comprehensive ground truth, applicable to different regions of the world, has been the major hindrance to the current study.

6 Future work

We plan to extensively validate our approach on more languages and countries with different characteristics. We also plan to extend the number and the nature of the considered ground truths. Since the ground truth data we used did not show adequate coherence nor completeness, further consideration on how to merge signals –also possibly coming from different domains– has to be put in place. An additional interest is to take into account signals extracted from social media by other automated tools, such as emergency-related classifiers, in order to introduce additional evidence. Also, data fusion techniques could be utilized to evaluate the effect of adding social media data to sensor or forecast data. A natural extension to the proposed network architecture could be to work with separate data sources and concatenate the results before the final dense layer.

Finally, an adaptive extension of the approach to dictionary building described in Section 3.2 is foreseen. This approach would be aimed at adaptively updating the dictionary while a newly detected event unfolds, thus enhancing both precision and recall of the subsequent data processing pipeline.

Acknowledgements The work at Politecnico di Milano and IIIA-CSIC was funded by the European Commission H2020 Project Crowd4SDG, #872944.

References

1. Autelitano, A., Pernici, B., Scalia, G.: Spatio-temporal mining of keywords for social media cross-social crawling of emergency events. GeoInformatica **23**(3), 425–447 (7 2019)
2. Avvenuti, M., Cimino, M.G.C.A., Cresci, S., Marchetti, A., Tesconi, M.: A framework for detecting unfolding emergencies using humans as sensors. Springerplus **5**, 43 (Jan 2016)
3. Bono, C., Pernici, B., Fernandez-Marquez, J.L., Shankar, A.R., Mülâyim, M.O., Nemni, E.: TriggerCit: Early Flood Alerting using Twitter and Geolocation–a comparison with alternative sources. In: Proc. ISCRAM 2022, Tarbes, France (May 2022)
4. de Bruijn, J.A., de Moel, H., Jongman, B., de Ruiter, M.C., Wagemaker, J., Aerts, J.C.J.H.: A global database of historic and real-time flood events based on social media. Sci Data **6**(1), 311 (Dec 2019)
5. Havas, C., Resch, B., Francalanci, C., Pernici, B., Scalia, G., Fernandez-Marquez, J.L., Van Achte, T., Zeug, G., Mondardini, M.R.R., Grandoni, D., Kirsch, B., Kalas, M., Lorini, V., Rüping, S.: E2mc: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. Sensors **17**(12) (2017)

6. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: Survey summary. In: Companion Proc. of the The Web Conference WWW'2018, Lyon, France. p. 507—511 (2018)
7. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. Data Mining and Knowledge Discovery **33**(4), 917–963 (Jul 2019)
8. Jongman, B., Wagemaker, J., Romero, B.R., De Perez, E.C.: Early flood detection for rapid humanitarian response: Harnessing near real-time satellite and twitter signals. ISPRS International Journal of Geo-Information **4**(4), 2246–2266 (2015)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014), cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015
10. Lindersson, S., Brandimarte, L., Mård, J., Di Baldassarre, G.: A review of freely accessible global datasets for the study of floods, droughts and their interactions with human societies. WIREs Water **7**(3), e1424 (2020)
11. Lorini, V., Castillo, C., Nappo, D., Dottori, F., Salamon, P.: Social media alerts can improve, but not replace hydrological models for forecasting floods. In: 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). pp. 351–356. IEEE Computer Society, Los Alamitos, CA, USA (Dec 2020)
12. Munawar, H.S., Hammad, A.W.A., Waller, S.T.: Remote sensing methods for flood prediction: A review. Sensors **22**(3) (2022)
13. Nemni, E., Bullock, J., Belabbes, S., Bromley, L.: Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery. Remote Sensing **12**(16) (2020)
14. Ofli, F., Qazi, U., Imran, M., Roch, J., Pennington, C., Banks, V., Bossu, R.: A real-time system for detecting landslide reports on social media using artificial intelligence. In: Di Noia, T., Ko, I.Y., Schedl, M., Ardito, C. (eds.) Web Engineering. pp. 49–65. Springer International Publishing, Cham (2022)
15. Olteanu, A., Castillo, C., Diaz, F., Vieweg, S.: Crisislex: A lexicon for collecting and filtering microblogged communications in crises. Proceedings of the International AAAI Conference on Web and Social Media **8**(1), 376–385 (5 2014)
16. Perera, D., Seidou, O., Agnihotri, J., Mohamed Rasmy, A.W., Smakhtin, V., Coulibaly, P., Mehmood, H.: Flood early warning systems: A review of benefits, challenges and prospects (08 2019)
17. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans. Knowl. Data Eng. **25**(4), 919–931 (2013)
18. Shoyama, K., Cui, Q., Hanashimaa, M., Sano, H., Usuda, Y.: Emergency flood detection using multiple information sources: Integrated analysis of natural hazard monitoring and social media data. Science of the Total Environment **767**(144371), 1–11 (2021)
19. Stollberg, B., De Groeve, T.: The use of social media within the global disaster alert and coordination system (GDACS). In: Proc. 21st Intl. WWW Conf. pp. 703–706 (2012)
20. United Nations Office for the Coordination of Humanitarian Affairs: Five essentials for the first 72 hours of disaster response. https://www.unocha.org/story/five-essentials-first-72-hours-disaster-response, accessed: 2022-07-15
21. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 1578–1585 (May 2017)