Extracting Large Scale Spatio-Temporal Descriptions from Social Media

Carlo Bono, Barbara Pernici

Politecnico di Milano, DEIB, Via Giuseppe Ponzio, 34, 20133 Milano, Italia

Abstract

The ability to track large-scale events as they happen is essential for understanding them and coordinating reactions in an appropriate and timely manner. This is true, for example, in emergency management and decision-making support, where the constraints on both quality and latency of the extracted information can be stringent. In some contexts, real-time and large-scale sensor data and forecasts may be available. We are exploring the hypothesis that this kind of data can be augmented with the ingestion of semi-structured data sources, like social media. Social media can diffuse valuable knowledge, such as direct witness or expert opinions, while their noisy nature makes them not trivial to manage. This knowledge can be used to complement and confirm other spatio-temporal descriptions of events, highlighting previously unseen or undervalued aspects. The critical aspects of this investigation, such as event sensing, multilingualism, selection of visual evidence, and geolocation, are currently being studied as a foundation for a unified spatio-temporal representation of multi-modal descriptions. The paper presents, together with an introduction on the topics, the work done so far on this line of research, also presenting case studies relevant to the posed challenges, focusing on emergencies caused by natural disasters.

Keywords

social media, spatio-temporal event description, information mining

1. Introduction

The nature of social networks implies an intrinsic distributional variability, since the topics on social networks usually reflect, to different extents, events taking place in the world. Some items transport relevant information about some aspect of reality, usually in the form of unstructured and semi-structured data. For example, information about a specific event could be made available using text in some language and/or images and videos, and further contextualized with tags and conversations, extended with external links, and enriched with comments and interactions. These different dimensions of social media posts contain information, explicitly or implicitly, possibly related to actual events. "Possibly" since separating relevant[1] and irrelevant information is usually cumbersome, both because of input data quality and volume. Since we aim at extracting descriptions of large-scale events along the dimensions of space and time, we mainly focus on two operational aspects: the ability to automatically detect whether a certain event is actually ongoing, and the ability to extract quantitative, actionable measures as the

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy

carlo.bono@polimi.it (C. Bono); barbara.pernici@polimi.it (B. Pernici)
0000-0002-5734-1274 (C. Bono); 0000-0002-2034-9774 (B. Pernici)

[1] Relevant as in "informative with respect to to some chosen event, kind of event or aspect of reality"

event unfolds, possibly in an adaptive fashion. The former is meant to be used as a trigger for preempting data collection tasks. The latter has to be compared and complemented with other layers, such as sensor or forecast data, in order to enhance the understanding of reality. Both aspects are affected by the variability of the social media data distribution over space and time, and both can exploit the same variability.

The remainder of this work is structured as follows. In the next session, the main challenges and topics in the literature are reviewed. Section 3 illustrates the investigations conducted so far on these topics. Section 4 combines these results in an application customized for flood events. Finally, Section 5 sketches future medium-term research directions.

2. Related work

The opportunities and challenges for using social media in emergency management have been trending topics over the last years. Large-scale emergencies are not trivial to tackle with automated and practical tools. We focus on emergencies since there is an essential value, as well as they pose complex questions to be studied. The papers [1, 2] highlight recent opportunities, challenges, applications and directions for the use of social media data for emergency management at large. Vertical studies for the use of social media for specific events have been performed, such as [3] for earthquake events and [4, 5] for the case of flood events. Techniques for automatic image filtering on social network data, mainly using deep learning, have been studied in the field of emergency and social sensing in [6, 7].

Methodologies and resources to analyze social media posts in multiple languages are also essential, as discussed in [8] for lexicon building, and in [9] for an adaptive, cross-source approach to social media crawling. How and to which extent social media can be integrated into production-grade emergency management systems has been studied in [10] and [11]. Some of the current work builds on past experience with emergencies, as documented in [12].

Although many studies explicitly address detection, filtering, geolocation and multilinguality support, some limitations exist. Production-level requirements are hardly met for many kind of events. Multi-modal approaches are sometimes investigated in literature, but not widely adopted and usually limited to one of the tasks. The same holds true for data fusion approaches. Our investigation aims at designing an end-to-end framework for automatically deriving event characterizations, taking advantage of multi-modal and multi-source approaches in order to solve these challenges. Benchmarking and comparison with existing solutions will be performed on publicly available datasets and self-produced ones.

3. Issues and methods

3.1. Event detection and adaptation

Monitoring regions of the world for classes of events is one of the possible entry points for social media data collection and processing. Some notable sources of information can dispatch alerts and forecasts, as it happens with GDACS[2] and GloFAS[3].

In this context, we study both timeliness and geographical precision of information derived from social media, in particular tweets, as we analyze it compared to information available in authoritative sources. Not all events are covered with the same timeliness, so for large scale events, we are using posterior validated data to understand to which extent social media can support already available data, such as forecast and sensor data, as a detection mechanism.

As a first step, the use of dictionaries of query terms for social sensing is being evaluated. For each of the dictionary terms, the time series of the term usage over a recent time window can be obtained. This signal is then used to estimate if an event of interest is ongoing or not. A first exploration of this methodology can be found in [13], in which we focused on searching Twitter in many possible languages starting from a limited set of seed search keywords. We are studying how to leverage available ground truth in order to automatically build language-tailored resources. The current focus is on building language-centered dictionaries with a data-centered approach[4], minimizing human intervention and taking advantage of the temporal correlation between words and events. Feature windows built with these dictionaries are then fed to a supervised classifier, such as a CNN, and evaluated in a leave-one-event-out fashion. Our preliminary assessment shows that a low precision (<50%) and a good recall (~80%) triggering system can be achieved, with almost no human intervention. This result is achieved with negligible processing effort, and a comfortable request resolution.[5]

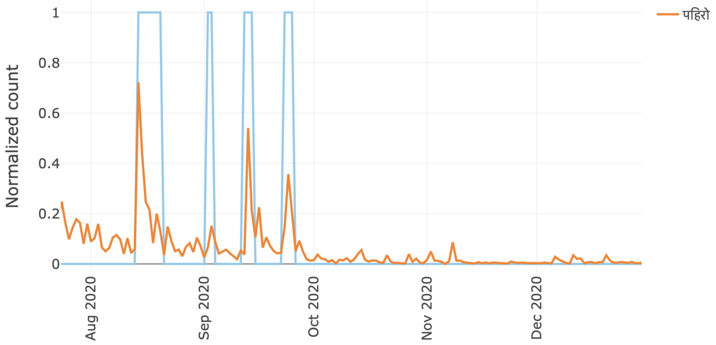


Figure 1: Occurrence of "landslide", compared with flood events in Nepal reported by GDACS

Detecting an event onset is only the first step towards awareness. Moreover, an efficient detection mechanism provides a precious byproduct: broad matching keywords related to events of interest. These, in principle, enhance the recall of the data collection. But as broad as they can be, two issues should be considered. On one side, unpredictable descriptors, event-specific qualities and linguistic variations are bound to happen. One significant example is given by location names, which are not usually constant across events. On the other side, good "general

[2] The Global Disaster Alert and Coordination System, started in 2004, is a joint initiative of the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) and the European Commission (EC).

[3] The Global Flood Awareness System is a component of the Copernicus Emergency Management Service (CEMS) that delivers global hydrological forecasts to registered users.

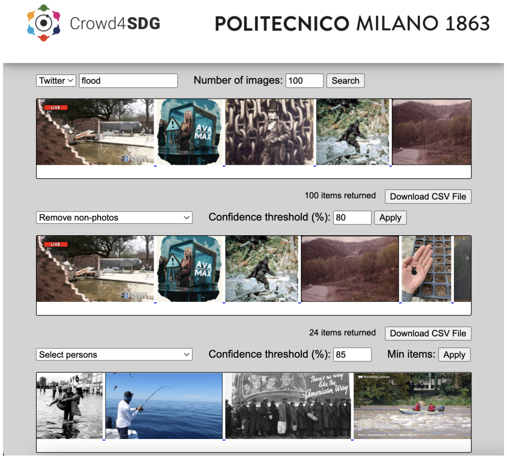
[4] Leveraging language models, word embeddings and external data sources

[5] One request per second https://developer.twitter.com/en/docs/twitter-api/rate-limits

purpose" predictive features do not guarantee to match relevant contents only. In light of this consideration, adaptive methods of data collection should be prioritized. Adaptability is aimed at controlling the specificity of an event, together with its evolution over space and time. An example of an approach that is both automated and adaptive is given by [9], focusing on identifying variable search keywords over time.

3.2. Understanding and filtering semi-structured contents

Methods and tools that extract structured information from social media are the main focus of this line of research. Among the general questions that can be posed, we include questions about the category of posts, as in "*is this post about* <x>*?*". These questions can be usually translated to supervised classification tasks. The questions being asked could be related to some specific semantic of the content, as in "*is there a* <y> *in this image?*". Some of these questions could still be mapped to classifications tasks, while some others relate to detection or segmentation tasks. A systematic approach for answering this kind of questions has been proposed in [7], developing the VisualCit[6] image processing toolkit, aimed at extracting indicators about emergency events. Here, image classification and object detection are usually achieved through the use of deep neural networks, both with off-the-shelf solutions and custom trained networks. In this way, specific queries can be posed to social media data, describing specific contents or situations, such as "two or more persons" or "a flood event". In addition, the ability to leverage user feedback to adjust classifiers is also a topic of interest. Pre-processing actions such as removing near-duplicates and filtering non-photographic material are also supported.



Twitter flood Number of images: 100

100 items returned

Remove non-photos Confidence threshold (%): 80

24 items returned

Select persons Confidence threshold (%): 85 Min items:

Figure 2: VisualCit interface for pipeline sketching

Text classification techniques, as well as a fused representation of textual, image, and metadata, are areas of interest, although so far they have been inspected only preliminarily. The primary goal would be to track known attributes of a situation, such as phases or topics, in order to characterize it. Since in this case the inputs are formulated in natural language, special

[6] http://visualcit.polimi.it:7778/

attention is being paid to multilingualism. Multilingualism can be supported with the use and the development of language-independent tools and representations, such as language-agnostic embeddings, or approaches that support a relevant set of languages.

Characterization of social media posts can be almost naturally attached to a temporal dimension. As a first approximation, since the extracted information augments the stream of published media, it is reasonable to assume that the information relates to near-realtime circumstances. A more pressing issue is the one about the spatial dimension of the contents. If the contents are not natively geolocated, as it usually happens[7], the location of the posts should be inferred. This is usually a harder problem and faceted, since the location of the author could differ from the location(s) related to the contents, or the hints for understanding the location could be flimsy. A comprehensive approach to social media geolocation has been proposed in the CIME system [14], which is currently being used by VisualCit. CIME disambiguates candidate locations using Nominatim[8]. The disambiguation algorithm considers candidate locations, their distances, and their ranks in the OSM administrative hierarchy. CIME was exploited in the E2mC project [12] to support rapid mapping activities within Copernicus Emergency Management Services. Extracted geolocations can be further filtered with known geometries (e.g., countries being monitored) and/or density-based approaches, also leveraging in-post and between-post relationships.

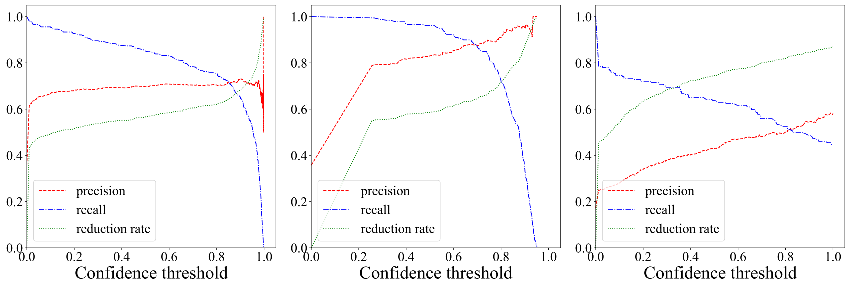
3.3. Building and enhancing pipelines

Comprehensive and quick reaction to large scale events requires a systematic approach to processing. Functionalities implemented so far are available with a service-based paradigm. Composite functionalities can be instantiated as a processing pipeline, using a straightforward configuration file, indicating components to be invoked and respective parameters, such as confidence thresholds. Of course, as it happens for the events themselves, the configuration of a pipeline is not always predictable. Data distributions, quality constraints and workforce availability are likely to change over time. With this fact in mind, we evaluated an iterative approach to provide suggestions to designers of data preprocessing pipelines. Iterations with user feedback are meant to rapidly achieve the desired goals, reflecting application needs and limitations and understanding the quality of the output. The designer is provided with an informative environment in which components, constraints, and parameters can be evaluated. Components typically consist of data filtering or augmentation modules, representing implementations of the tasks introduced in 3.2. Constraints are usually expressed on processing quality, efficiency and cost. Figure 3 shows an example dashboard in which the effect of the pipeline components on the final output is visualized interactively. The effect is measured in terms of precision, recall and number of filtered items. The evaluation is performed using a validated sample dataset, provided during the design process.

Processing posts with tailored pipelines dramatically reduces the number of posts. Depending on application needs, the number of items that are filtered out and the quality of the output can be balanced. Suggestions about possible improvements are provided, such as modifications to the configuration of the pipeline, based on historical or current evidence. Suggestions could

[7] For example, for privacy issues.

[8] https://nominatim.openstreetmap.org/



(a) Is this a photo?

(b) Are there two or more persons?

(c) Is this a public place?

Figure 3: Precision, recall and reduction rate responses to confidence threshold, on an example pipeline

also concern the execution order of the pipeline, for performance optimization. We showed that, in some test cases, an optimized configuration could lead to around 33% of the original execution time, without affecting the output.

4. Case studies

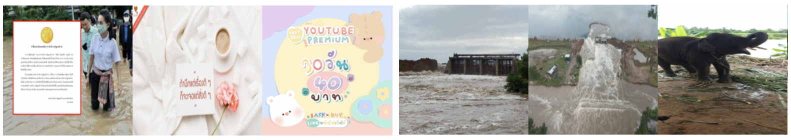
As case studies, we exploited, extended and adapted available tools to assess their reliability when used as a triggering and analysis system for large scale disaster events. We applied a combination of the proposed methodologies to two flood events that occurred in 2021 in Thailand (September) and Nepal (June and July). The two events were selected since the United Nations Satellite Centre (UNOSAT) supported both activations in Nepal and Thailand with satellite-derived maps and AI-based flood detection analysis. In both case studies, we used Twitter as a data source.

Initial evidence suggested the need for language-specific approaches rather than focusing on contents written in English. Regarding event detection, the experimentation was started with a limited amount of words. The word count signal over time proved to be sufficient to detect the two events within a reasonable time frame. Using small dictionaries is less sensitive to noise but poses a number of limitations, which have been discussed in subsection 3.1. A pipeline consisting of the following tasks has been run in order to select relevant data:

1. Remove duplicated images and similar images.
2. Remove non-photos, such as drawings, screenshots and computer-generated images.
3. Remove not-safe-for-work images[9].
4. Geolocate posts using CIME.

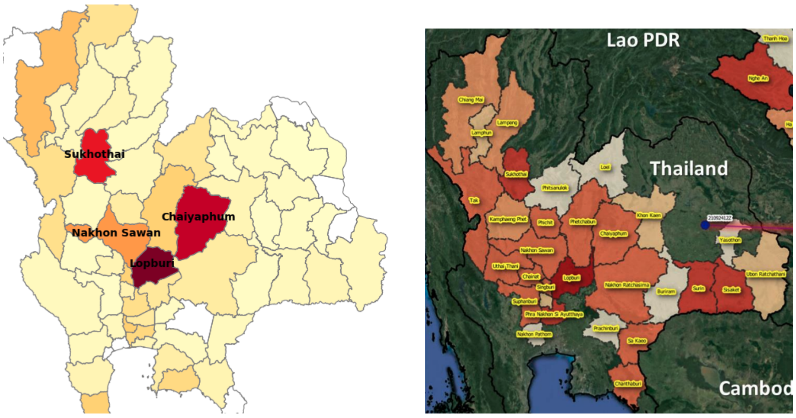
Processing the posts with such a pipeline reduced the amount of data by orders of magnitude. The output, aggregated by administrative region, was compared to official impact estimates (Fig. 4 and 5) shows promising yet improvable results. We were able to process and geolocate images at a rate of roughly 10,000 items per hour, on a single server machine. This proved to be sufficient for real-time event monitoring.

[9] "*Not suitable for viewing at most places of employment*", according to the Merriam-Webster online dictionary.



(a) Removed items (b) Kept items

Figure 4: Example images for (a) items being removed by the pipeline, and (b) items being kept



(a) Geolocations / inhabitants ratio by region, normalized by population

(b) Affected persons by region at September, 28th (source: ReliefWeb)

Figure 5: Comparing pipeline output and preliminary impact assessment, Thailand case study [13]

5. Future work

The current work in progress is aimed at detection methods, components for data extraction, and process design support. Together, these elements can be leveraged to project the streams of raw data onto spatio-temporal coordinates, in order to get a suitable and actionable description of reality. So far, official, sensor and forecast data have been used as a reference for evaluating the quality of the approaches. Also, they could be fused together, in order to build a multi-faceted view of reality. This is the guiding goal for future work on the topic. At the same time, since social media contents are highly multi-modal by definition, we want to explore how different dimensions and aggregations of social media data can be combined in multi-modal representations in order to better fit the required functionalities, such as classification tasks.

Acknowledgements This work was funded by the EU H2020 project Crowd4SDG “Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience”, #872944.

References

[1] M. Imran, F. Ofli, D. Caragea, A. Torralba, Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions,

Information Processing & Management 57 (2020) 102261.

[2] V. Lorini, C. Castillo, S. Peterson, P. Rufolo, H. Purohit, D. Pajarito, J. P. de Albuquerque, C. Buntain, Social media for emergency management: Opportunities and challenges at the intersection of research and practice, in: 18th International Conference on Information Systems for Crisis Response and Management, 2021, pp. 772–777.

[3] T. Sakaki, M. Okazaki, Y. Matsuo, Tweet analysis for real-time event detection and earthquake reporting system development, IEEE TKDE 25 (2013) 919–931.

[4] J. Fohringer, D. Dransch, H. Kreibich, K. Schröter, Social media as an information source for rapid flood inundation mapping, Natural Hazards and Earth System Sciences 15 (2015) 2725–2738.

[5] K. Shoyama, Q. Cui, M. Hanashimaa, H. Sano, Y. Usuda, Emergency flood detection using multiple information sources: Integrated analysis of natural hazard monitoring and social media data, Science of the Total Environment 767 (2021) 1–11.

[6] D. T. Nguyen, F. Alam, F. Ofli, M. Imran, Automatic image filtering on social networks using deep learning and perceptual hashing during crises, in: Proc. 14th ISCRAM Conf. – Albi, France, 2017.

[7] V. Negri, D. Scuratti, S. Agresti, D. Rooein, G. Scalia, A. R. Shankar, J. L. Fernandez-Marquez, M. J. Carman, B. Pernici, Image-based social sensing: Combining AI and the crowd to mine policy-adherence indicators from twitter, in: 43rd IEEE/ACM Intl. Conf. on Sw. Eng. Track Software Engineering in Society, ICSE (SEIS), Madrid, Spain, IEEE, 2021, pp. 92–101.

[8] A. Olteanu, C. Castillo, F. Diaz, S. Vieweg, Crisislex: A lexicon for collecting and filtering microblogged communications in crises, Proceedings of the International AAAI Conference on Web and Social Media 8 (2014) 376–385.

[9] A. Autelitano, B. Pernici, G. Scalia, Spatio-temporal mining of keywords for social media cross-social crawling of emergency events, GeoInformatica 23 (2019) 425–447.

[10] B. Stollberg, T. De Groeve, The use of social media within the global disaster alert and coordination system (GDACS), in: Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 703–706.

[11] V. Lorini, C. Castillo, D. Nappo, F. Dottori, P. Salamon, Social media alerts can improve, but not replace hydrological models for forecasting floods, in: 2020 IEEE/WIC/ACM Intl. Joint Conf. on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, 2020, pp. 351–356.

[12] C. Havas, B. Resch, C. Francalanci, B. Pernici, G. Scalia, J. L. Fernandez-Marquez, T. Van Achte, G. Zeug, M. R. R. Mondardini, D. Grandoni, et al., E2mC: Improving emergency management service practice through social media and crowdsourcing analysis in near real time, Sensors 17 (2017) 2766.

[13] C. Bono, B. Pernici, J. L. Fernandez-Marquez, A. R. Shankar, M. O. Mülâyim, E. Nemni, Triggercit: Early flood alerting using twitter and geolocation–a comparison with alternative sources, arXiv preprint arXiv:2202.12014, accepted for presentation at ISCRAM'22 (2022).

[14] G. Scalia, C. Francalanci, B. Pernici, Cime: Context-aware geolocation of emergency-related posts, GeoInformatica (2021) 1–33.