Adaptive Filtering Strategies for Social Media Streams

Carlo A. Bono[1],*

[1]Politecnico di Milano, DEIB Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Abstract

Information produced on social media platforms can serve a wide range of applications, yet the ability to obtain and refine this information is hindered. Social media data naturally exhibit a multi-modal and relational nature, with challenges associated with its volume, pace, variability, and a variety of sources of noise. Especially in time-sensitive scenarios, these characteristics call for efficient, automated filtering approaches capable of extracting the available relevant data with minimal human supervision. How these automated approaches can be optimized while respecting context-dependent quality constraints is an understudied issue. Moreover, adaptation to non-stationarity in both data and relevance enriches the perspective of this research topic. The study of a method for overcoming these challenges by leveraging an adaptive architecture is proposed. Adaptivity pertains to selecting data representations, their aggregation, and the filtering decision policy. These choices are subject to operating constraints over the quality dimensions of accuracy, completeness and timeliness. The research question is contextualized in the state of the art, and its novel aspects are discussed. Preliminary results are described, together with a research plan outline.

Keywords

Social Media Data Quality, Adaptive systems, Multi-modal data filtering

> Je ne suis heureux que lorsque j'ai trouvé
> une «formule»
>
> *Cahiers 1957-1972*
> *Emil Cioran*

1. Research goal formulation

The research objective is an adaptive method for filtering relevant content from social media streams, as defined by a context-dependent desired goal. This method is the goal artifact, and its design is the top-level practical problem guiding the exploratory research effort.

Contents on social media and their relations are ever-changing, reflecting the unfolding of events in the real world. Extracting relevant information from social media poses specific challenges due to the volume and pace of data, the sensitivity of data filtering to time, contexts and constraints, the multi-modal nature of contents, the dynamic relationships that emerge

CAiSE '23: Doctoral Consortium co-located with the 35th International Conference on Advanced Information Systems Engineering, CAiSE '23, June 12-16, 2023, Zaragoza, Spain

*Corresponding author.

carlo.bono@polimi.it (C. A. Bono)
0000-0002-5734-1274 (C. A. Bono)

among them, and the difficult trade-off between quality measures.

The study pertains to the early stages of data preparation pipelines, intersecting data *discovery* and *cleansing*. The method is intended to balance data quality constraints in terms of relevance, completeness and timeliness. The knowledge questions that stem from the driving objective are related to approaches and algorithms functional to two main properties of the method under study: relevance of the outputs and timely adaptiveness. *Relevance* is to be studied as a function that emerges from content representations, relations among contents, and user-defined goals and constraints. *Adaptiveness* is to be investigated by taking into account changing data distributions and relations.

The end users of the method would be actors interested in a fast, machine-assisted assembly of data analysis pipelines for social media streams, with an accent on the adaptive selection of significant results. This is motivated by scenarios in which events unfolding over time, possibly abruptly or unpredictably, have to be monitored. Decisions about collection and filtering often translate to time-consuming experiments, burdensome procedures and single-use implementations, whose performance degrades over time. In several scenarios, the need to rapidly and adaptively extract meaningful data dominates other resource constraints, while still requiring a governable quality of extracted data. The consumers of the outputs of the applications would be designers of enclosing data preparation pipelines, data analysts, or downstream tasks such as supervised classification tasks.

Example use cases for social media analysis tasks are business purposes, political communications and emergency management. In all these cases, valuable information lies among a sheer amount of posts, including comments, opinions, machine-generated content, and unrelated messages. Data could be intentionally tainted by misinformation or spam. Social media data share many of the characteristics of the so-called big data. Tackling the distinctive volume, velocity, and variability of social media calls for automated approaches that focus on different aspects of data understanding in order to control the quality of the output. In many scenarios, data quality is controlled by combining these approaches and human expertise to validate the reliability of the information extracted. We argue that this expertise can be wired into a closed-loop paradigm involving both human and automated operations in order to build and update a data processing policy aimed at maximizing the relevance of the output while controlling both its timeliness and completeness.

2. Related work

Through a literature analysis, the authors of [1] highlight a general lack of research on the stages of data discovery, collection, and preparation in social media analytics. They report that data volume is mainly cited as a challenge in the literature, while other categories have received less attention. An overview of the predominant research directions in the field can be found in [2] and [3], underlining active research fields in social media analytics ranging from marketing to information sharing during emergencies to politics. For one of the relevant application fields of social media analysis, disaster management, [4] reports that studies involving the dimensions of time, content and network together are underrepresented. Coherently with [1], also in [5] the technology-related data quality issues are reviewed with the perspective of the big data “five Vs”. Their work underlines the ongoing debate about the quality of social media data, questioning whether it is relevant for generalization in the context of research and development. The framework proposed in [6] is also grounded in a similar perspective, but through the lens of service composition, aiming at providing a quality model to capture the mutating features of social media data. Authors in [7] analyze readability, completeness, usefulness, and trustworthiness in the context of a social media platform, Twitter. Another social media quality framework for Twitter is proposed in [8]. Challenges and approaches specific to data cleaning are reviewed in [9]. [10] proposes a reinforcement learning technique for selecting an optimal sequence of data processing steps with a given dataset and quality performance metric, emphasizing that the study of automatically deriving optimal data pre-processing pipelines has been understudied. [11] aims at supporting non-experts with the estimation of the impact of pre-processing operators in a machine learning scenario.

Among the above-mentioned application domains, emergency management poses interesting constraints both to the required quality and the timeliness of the operations. Social media can provide first-hand information under tight time constraints for the analysis of emergencies as described, for instance, in [12] for earthquake reporting. Adaptation in quantitative analysis related to trending topics during COVID-19 is analyzed in [13]. [14] uses frequency features of tweets to infer flood events on a global scale. [15] proposes a framework to detect composite social events over social media streams.

Despite the huge amount of works leveraging text information extracted from social media platforms, in many scenarios the analysis of multimedia content is of paramount importance. A notable scenario is again emergency management, where the presence of an appropriate image is a strong proxy of relevance [16]. In recent years deep learning approaches leveraging the integration between image and text data were presented, such as [17]. A recent survey on the state of the art of representation learning techniques is presented in [18]. Embedding models can be used to extract representations of both documents and users. In [19] and [20], a method for creating semantically meaningful representations of users is presented. Higher-level entities can also be investigated. Among analytical studies to the study of communities, polarization and echo chambers, [21] proposes a method that highlights communities of users sharing a stance in the COVID-19 debate, examining the structure of the networks and the textual content of the interactions. [22] proposes a scalable model to estimate the political leanings of Twitter users, leveraging network structure and users’ profile descriptions.

Social media expose vast amounts of its users to information campaigns, visible or malign, that influence collective opinions. Heuristics for representation methods aiming at supporting veracity can be derived from misinformation analysis techniques. Technical challenges for discriminating trends among users are pinpointed in a machine learning framework proposed in [23]. The authors focus on the challenge of early detection of promoted content, highlighting that content, user, network and timing features play different roles with respect to the evolution of the events. [24] analyzes the temporal dynamics of misinformation spreaders in a dynamic graph-based framework, proposing a detection approach. [25] proposes a semi-supervised approach for identifying bots that learns a joint representation of social connections and interactions among users, leveraging graph-based representation learning and label propagation. [26] studies the spread of propaganda and misinformation that circulated during the first months of the Russia-Ukraine conflict, concluding that Facebook and Twitter are vulnerable to abuse during crises.

Machine learning methods applied to graphs have been an active research area in the last few years. [27] aims at learning low-dimensional, continuous feature representations for nodes in a network. [28] proposes a semi-supervised convolutional learning method for graphs. The method scales linearly in the number of graph edges and encodes both local graph structure and node features. [29] surveys representation learning literature in the context of dynamically evolving graphs. Studies on the evolution of graphs are underrepresented in graph machine learning. An example is given in [30], presenting a learning framework for dynamic graphs.

3. Challenges and research questions

Literature review and exploratory work highlighted an assortment of challenges that can be translated into knowledge questions. These questions focus on the desired properties of the method under study: relevance of the produced outputs and adaptivity, as defined by the combination of application goals and suitable quality measures.

3.1. Relevance in a non-stationary environment

Social media data show remarkable characteristics of variety and variability. These characteristics evolve over time in reaction to events. If user feedback about the relevance of data is produced, suitable approaches can be derived in order to connect the feedback to the extraction process. This feedback enables the method to isolate relevant contents in an adaptive fashion. We assume that a specific application goal can be implicitly approximated by user feedback on the processing outputs. Several approaches in machine learning and network analysis model relevancy as an attribute linked to intensional or structural data properties, possibly accounting for data labels:

$$\begin{aligned} R &= f(data,\, labels^{*}) \\ R &= f(structure,\, labels^{*}) \end{aligned} \tag{1}$$

Building up on such techniques, our research question investigates how to assess the relevance of social media data from its multi-dimensional characteristics, network-derived features, and their evolution over time:

$$R = f(g(data),\, h(structure),\, time,\, labels) \tag{2}$$

With the g and h notations, we stress that the research objective is not directly related to the techniques of representation learning but rather to the interplay that relevance has with these constituents and their composition. Relevance is goal-dependent, and it is assumed to be time-dependent. The overall research question is formulated as follows:

- How to design an adaptive method for filtering relevant data from multi-modal, non-stationary social media streams?

3.2. Representations for data modalities and relations

Social media data naturally exhibit a multi-modal nature. Structured and unstructured content can be encoded in representations that contain valuable information for isolating relevant contents. Social media data also exhibit a relational nature since entities are linked among them, forming a network, and non-atomic entities emerge from these structures, such as authors, threads and communities. The nature and the impact of the representations for data and their relations are key aspects of the proposed research. A visual intuition is proposed in Figure 1. We are interested in the following knowledge questions stemming from the main research question:

- What are the trade-offs between the choices on representations extracted from social media data, operation-level constraints and efficacy of the adaptation?
- How the representations of the different dimensions can be combined or aggregated? Are there paradigmatic changes that depend on the scenario?

3.3. Time dimension

We assume that application constraints are specified by an application designer and explicitly or implicitly adjusted at operation time. We also assume that human feedback on the relevancy of the outputs will be available, at least for a subset of items and with some latency. This feedback could again be explicit, at the design phase or implicit, in subsequent steps of the data processing or use. We are interested in exploring the following knowledge questions that relate to the timeliness dimension:

- What are the trade-offs among latency in the *delivery* of the results, latency in the *adaptation* to the feedback, and latency in the *design* of application instances?
- Are there approaches that are sub-optimal in terms of relevance and yet can be applied under specific timeliness constraints?

4. Contributions

The contribution is focused on supporting the discovery and filtering phases with an automated adaptive approach for social media data, which is a germinal data preparation topic. Which algorithms and techniques are better suited to select transformation actions, their orchestration and aggregation, and under which conditions and scenarios, are understudied knowledge questions according to the literature. These questions are tackled mainly using quantitative experimental approaches over real-world data. A supplementary contribution is the joint study of the target adaptive methods and the dimension of data timeliness, enriching the assessment of the method evaluation with faceted performance dimensions.

An overview of the reference framework for investigating the research question is depicted in Figure 2, describing logical components, data objects, and data flows. The subject of the adaptation is a data selection policy. The policy defines an extraction and filtering process. The driver of adaptation is user feedback as a proxy measure of relevance. Process constraints are

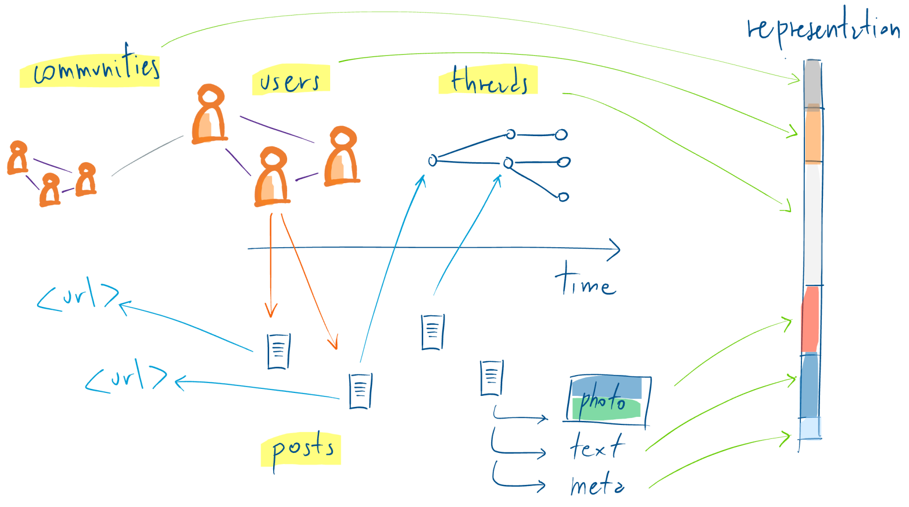

Figure 1: Data structures, relations and their representations on social media

expressed in terms of input and output throughput, output accuracy, and output timeliness. The data selection policy is partitioned into sub-policies related to functional areas. The combined effect of the sub-policies expresses the overall policy at time t, which is denoted as P_t.

Data streams originating from one or more social media platforms are queried based on a *query policy*, $P_{(Q,t)}$. Intra-entity and inter-entity characteristics of the extracted data are mapped to representations leveraging a transformation action library A. Actions can apply to individual data items or a set of data items. The choice of which actions to apply depends on a *transformation policy*, $P_{(T,t)}$. Representations for data items are then aggregated, and an estimate of the relevance is produced by a *filtering policy*, $P_{(F,t)}$.

User feedback is produced at design time or along the downstream processing. The feedback could be on data relevancy, timeliness, or both. It is assumed that the full or a restricted set of feedbacks can be obtained and stored. Each feedback is related to a uniquely identified item. It is assumed that data coming from the original streams cannot be stored due to restrictions on the storage volume and on the retention policy. It is assumed that a restricted set of transformed representations, which are anonymized, can be stored. For this purpose, a representation store S_R is provided. The representation store could reference the original content, which is volatile, and traces the $P_{(T,t)}$ transformation policy. Representations and their selection and aggregation are time-dependent. Policies are tracked and historicized.

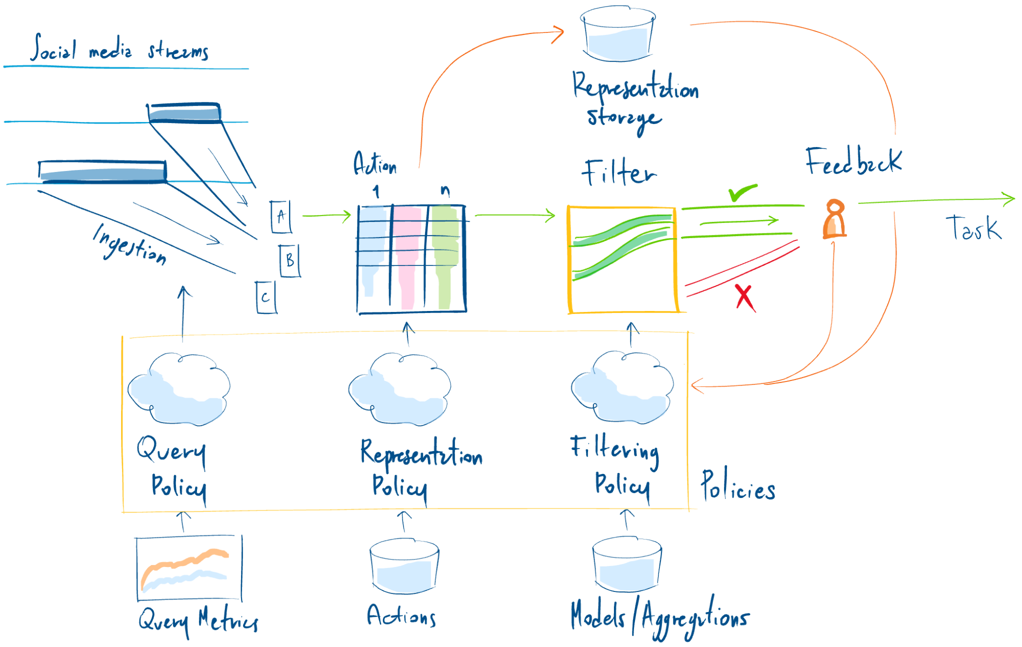

Figure 2: Reference framework for the proposed investigation

5. Preliminary results and planning

5.1. First year

During the first year, the study has been focused on data preparation pipelines tailored to social media. A conceptual framework for machine-assisted, human-in-the-loop pipeline design has been elaborated in [31]. Case studies for pipeline design in emergency management scenarios were presented in [32]. The approach has been enriched with a dictionary-based adaptive data ingestion and detection method in [33]. Both approaches leverage textual and image data in order to identify the onset of natural disaster events and extract relevant contents during the first hours of emergency response. A contextualization of the method in the more general framework of analyzing social media with crowdsourcing can be found in [34].

5.2. Second year

The main line of work for the second year is a comprehensive evaluation of relevant data representations, combination strategies, and their impact on relevance. As an experimental scenario, an investigation on how to assess the severity of cybersecurity threats based on social media content and structure is in progress. In parallel, the impact of mixed representation methods on the detection of misinformation in social media is being studied. The late part of the second year will be devolved to the assessment of algorithms for deriving the optimal filtering policy $P_{(F,t)}$ depending on application constraints, together with the exploration of performance trade-offs, especially depending on representation choices. These evaluations are aimed at tackling the questions specified in subsection 3.2.

5.3. Third year

The third year will be devolved to the study of the implications of the policies and their combinations on the time dimension, focusing on the topics raised in subsection 3.3 and experimenting on the different sources of latency (design, processing, adaptation) and their interplay. Appropriate validation scenarios, with real-time or simulated dynamics using relevant datasets, will be examined. It is also expected that the study of filtering policies initiated during the second year will continue throughout the third year.

Acknowledgments

I thank Professor Barbara Pernici, my research supervisor, for her patient support and vision. I thank all the colleagues at the Information Systems group at Politecnico di Milano for their help and advice, and the colleagues in the Crowd4SDG project for the fruitful, amicable collaboration. This work was funded by the European Commission H2020 project Crowd4SDG "Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience", #872944.

References

[1] S. Stieglitz, M. Mirbabaie, B. Ross, C. Neuberger, Social media analytics–challenges in topic discovery, data collection, and data preparation, International journal of information management 39 (2018) 156–168.

[2] C. Zachlod, O. Samuel, A. Ochsner, S. Werthmüller, Analytics of social media data–state of characteristics and application, Journal of Business Research 144 (2022) 1064–1076.

[3] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, S. Nerur, Advances in social media research: Past, present and future, Information Systems Frontiers 20 (2018) 531–558.

[4] Z. Wang, X. Ye, Social media analytics for natural disaster management, International Journal of Geographical Information Science 32 (2018) 49–72.

[5] A. K. Srivastava, R. Mishra, Analyzing social media research: A data quality and research reproducibility perspective, IIM Kozhikode Society & Management Review 12 (2023) 39–49.

[6] K. Ali, M. Hamilton, C. Thevathayan, X. Zhang, Big social data as a service (bsdaas): a service composition framework for social media analysis, Journal of Big Data 9 (2022) 64.

[7] F. Arolfo, K. C. Rodriguez, A. Vaisman, Analyzing the quality of twitter data streams, Information Systems Frontiers (2022) 1–21.

[8] C. Salvatore, S. Biffignandi, A. Bianchi, Social media and twitter data quality for new social indicators, Social Indicators Research 156 (2021) 601–630.

[9] X. Chu, I. F. Ilyas, S. Krishnan, J. Wang, Data cleaning: Overview and emerging challenges, in: Proceedings of the 2016 international conference on management of data, 2016, pp. 2201–2206.

[10] L. Berti-Equille, Active reinforcement learning for data preparation: Learn2clean with human-in-the-loop., in: 10th Annual Conference on Innovative Data Systems Research (CIDR '20), 2020.

[11] B. Bilalli, A. Abelló, T. Aluja-Banet, R. Wrembel, Presistant: Learning based assistant for data pre-processing, Data & Knowledge Engineering 123 (2019) 101727.

[12] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th international conference on World wide web, 2010, pp. 851–860.

[13] 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 International Conference on Software Engineering: Software Engineering in Society, ICSE (SEIS) 2021, Madrid, Spain, May 25-28, 2021, IEEE, 2021, pp. 92–101.

[14] J. A. de Bruijn, H. de Moel, B. Jongman, M. C. de Ruiter, J. Wagemaker, J. C. J. H. Aerts, A global database of historic and real-time flood events based on social media, Sci Data 6 (2019) 311.

[15] X. Zhou, L. Chen, Event detection over twitter social media streams, The VLDB journal 23 (2014) 381–400.

[16] R. Peters, J. P. de Albuquerque, Investigating images as indicators for relevant social media messages in disaster management., in: ISCRAM, 2015.

[17] X. Huang, Z. Li, C. Wang, H. Ning, Identifying disaster related social media for rapid response: a visual-textual fused cnn architecture, International Journal of Digital Earth 13 (2020) 1017–1039.

[18] L. Ericsson, H. Gouk, C. C. Loy, T. M. Hospedales, Self-supervised representation learning: Introduction, advances, and challenges, IEEE Signal Processing Magazine 39 (2022) 42–62.

[19] I. R. Hallac, S. Makinist, B. Ay, G. Aydin, user2vec: Social media user representation based on distributed document embeddings, in: 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), IEEE, 2019, pp. 1–5.

[20] D. Irani, A. Wrat, S. Amir, Early detection of online hate speech spreaders with learned user representations., in: CLEF (Working Notes), 2021, pp. 2004–2010.

[21] G. Crupi, Y. Mejova, M. Tizzani, D. Paolotti, A. Panisson, Echoes through time: Evolution of the italian covid-19 vaccination debate, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 16, 2022, pp. 102–113.

[22] J. Jiang, X. Ren, E. Ferrara, Retweet-bert: Political leaning detection using language features and information diffusion on social networks, arXiv preprint arXiv:2207.08349 (2022).

[23] O. Varol, E. Ferrara, F. Menczer, A. Flammini, Early detection of promoted campaigns on social media, EPJ Data Science 6 (2017).

[24] J. Plepi, F. Sakketou, H.-J. Geiss, L. Flek, Temporal graph analysis of misinformation spreaders in social media, in: Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 89–104.

[25] M. Mendoza, M. Tesconi, S. Cresci, Bots in social and interaction networks: detection and impact estimation, ACM Transactions on Information Systems (TOIS) 39 (2020) 1–32.

[26] F. Pierri, L. Luceri, N. Jindal, E. Ferrara, Propaganda and misinformation on facebook and

twitter during the russian invasion of ukraine, in: 15th ACM Web Science Conference 2023, 2023.

[27] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–864.

[28] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907 (2016).

[29] S. M. Kazemi, R. Goel, K. Jain, I. Kobyzev, A. Sethi, P. Forsyth, P. Poupart, Representation learning for dynamic graphs: A survey, The Journal of Machine Learning Research 21 (2020) 2648–2720.

[30] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, M. Bronstein, Temporal graph networks for deep learning on dynamic graphs, arXiv preprint arXiv:2006.10637 (2020).

[31] C. A. Bono, C. Cappiello, B. Pernici, E. Ramalli, M. Vitali, A conceptual model for data analysis pipelines: Supporting the designer in datasets construction, submitted to Journal of Data and Information Quality (2022).

[32] C. A. Bono, B. Pernici, J. L. Fernandez-Marquez, A. R. Shankar, M. O. Mülâyim, N. Edoardo, et al., Triggercit: Early flood alerting using twitter and geolocation-a comparison with alternative sources, in: 19th International Conference on Information Systems for Crisis Responseand Management,{ISCRAM} 2022, Tarbes, France, May 22-25, 2022, ISCRAM Digital Library, 2022, pp. 674–686.

[33] C. A. Bono, M. O. Mülâyim, B. Pernici, Learning early detection of emergencies from word usage patterns on social media, in: 7th IFIP WG5.15 Information Technology in Disaster Risk Reduction, 2022, Kristiansand, Norway, October 12-14, 2022.

[34] C. Bono, M. O. Mülâyım, C. Cappiello, M. Carman, J. Cerquides, J. L. Fernandez-Marquez, R. Mondardini, E. Ramalli, B. Pernici, A citizen science approach for analysing social media with crowdsourcing, IEEE Access (2023).