

Leveraging visual analytics of Volunteered Geographic Information to support impact-based weather warning systems

Carlo Navarra^a, Kostiantyn Kucher^b, Tina-Simone Neset^a,
Clara Greve Villaro^c, Fredrik Schück^c, Jonas Unger^b, Katerina Vrotsou^b

^a Department of Thematic Studies–Environmental Change, Centre for Climate Science and Policy Research, Linköping University, Linköping, 58183, Sweden

^b Department of Science and Technology, Division for Media and Information Technology, Linköping University, Campus Norrköping, Norrköping, 60174, Sweden

^c Swedish Meteorological and Hydrological Institute, Norrköping, 60176, Sweden

ARTICLE INFO

Keywords:

Visualization
Classification
Volunteered Geographic Information (VGI)
Social media data
Extreme weather events
Flooding

ABSTRACT

As extreme weather events such as floods, storms, and heatwaves proliferate, local and regional authorities face challenges in predicting, monitoring, and assessing these events and their impacts. The introduction of impact-based warning services requires detailed, location-specific information on local vulnerability and impacts. This necessitates complementing conventional data with insights from local actors, and to explore novel methods for relevant public data monitoring through social media and news outlets. This paper presents a visual analytics pipeline that was co-developed with practitioners, aiming to detect impacts of extreme weather events, particularly floods, using Volunteered Geographic Information (VGI). The pipeline steps include: collecting VGI from social media, classifying and analysing the data, and visualizing it through an interactive interface. An empirical evaluation study was performed with meteorological and hydrological experts to assess the developed visual interface. The study collected and analysed feedback on the usability of the interface and identified interaction patterns from the experiment's screen recordings.

1. Introduction

As the world is facing a rise in extreme weather events, such as floods, storms, and heatwaves [1], local and regional authorities face increased challenges in terms of prediction and monitoring of such events, in assessing the risk of their impacts, and implementing appropriate adaptation measures [2]. While meteorological and hydrological agencies specialize in forecasting severe weather events and issue warnings, anticipating the local impacts of such events presents a significant challenge. To bridge this gap, following the guidelines of the World Meteorological Organization, national agencies worldwide have been introducing impact-based weather warning services [3].

Extending from traditional forecasting approaches based on meteorological and hydrological models to impact-based ones implies shifting focus from what the weather will be to what the weather will do [4, page 12]. Such a shift requires better insight into local vulnerabilities and local impacts, on population and infrastructure, of severe weather events which increases demands for current, location-specific and high-resolution information down to the individual, activity or community level [4]. To achieve this, conventional information sources, such as weather stations, weather radars, and satellites, are complemented in consultation

* Corresponding author.

E-mail address: carlo.navarra@liu.se (C. Navarra).

<https://doi.org/10.1016/j.ijdr.2025.105562>

Received 20 February 2025; Received in revised form 5 May 2025; Accepted 8 May 2025

Available online 27 May 2025

2212-4209/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

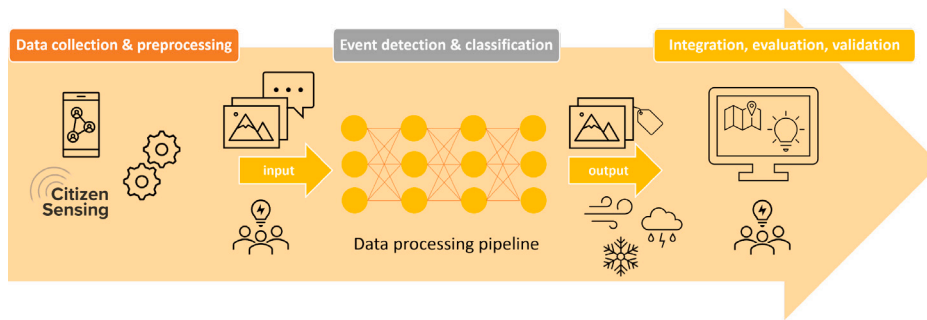


Fig. 1. VA pipeline developed to support impact-based weather warning systems.

with local, municipal, and regional actors to capture local characteristics, level of exposure and variations of possible impacts. Increasingly, such sources are also further informed through external monitoring of news outlets and social media, which could be particularly relevant for urban environments, where the frequency of social media posts is higher.

Motivated by the need for increased understanding of the characteristics and extent of local impacts, such as damage to infrastructure, property and to the environment, this paper presents a Visual Analytics (VA) pipeline for algorithmically detecting and visually investigating the space, time, and contextual dimensions of flooding events and their impacts from opportunistic Volunteered Geographic Information (VGI) [5] (Fig. 1). Opportunistic VGI refers to information sources such as social media channels where the intention of the contributing individual is not to share explicit observations of a flooding event. Nevertheless relevant information can be extracted from the information posts. VGI in the form of social media are of particular interest because of their ubiquitous nature compared to participatory VGI approaches requiring explicit intentional observations submitted through dedicated websites or apps. The analysis and visualization of opportunistic VGI can thus provide access to an extensive data set that captures and reflects various types of observation and their context, offering insights into the geographic extent, severity and temporality of extreme weather events. As such, it has potential practical implications for weather warning services but also for strategic planning and climate adaptation.

The VA pipeline presented here was co-developed as part of the AI4ClimateAdaptation (AI4CA)¹ research project and ran alongside the launch of the Swedish national system for impact-based weather warnings in 2021. The development of the VA pipeline was based on a collaborative process with practitioners involved in the national and regional processes of the impact-based weather warning system [6]. Initial experimentations with basic algorithms for classifying only text data were presented in Styve et al. [7], leading to the subsequent formulation of key aspects and potential pathways for the design and development of such a VA pipeline as discussed in Vrotsou et al. [8]. In this paper, we present the design and evaluation of the VA pipeline which comprises three key stages: (1) automatic retrieval of text and image data from social media, (2) Machine Learning (ML)-based classification of these texts and images as relevant/non-relevant to extreme weather events, and (3) visualization and exploration of the classified data. This study makes three main contributions:

- (i) the introduction of an automated approach for accessing relevant information to support impact-based warnings through ML-based classification of multimodal VGI,
- (ii) a visualization interface for the multifaceted exploration of VGI of extreme weather events in relation to weather warnings, and
- (iii) an empirical evaluation with weather warning practitioners, focusing on the relevance and usability of the pipeline by means of interaction pattern analysis and an analysis of semi-structured interviews.

The proposed VA pipeline offers several key advantages. Classification of VGI using transformer-based ML techniques allows for rapid categorization and filtering of relevant information, reducing noise and improving accuracy. Visualization through an interactive interface provides domain experts with a tool to explore the impacts of extreme weather events in the spatio-temporal context in which they occur. Additionally, the use of spatio-temporal visualization addresses inherent biases and limitations in VGI, such as unequal spatial and temporal distribution, ensuring more accurate and reliable analysis. The combination of these methodologies creates a system that has the potential to bridge the gap between raw social media data and actionable insights for impact-based weather warning services.

The remainder of this article is structured as follows. *Related work* is discussed first in Section 2. Section 3 outlines materials and methods used in this work. This includes the co-design process followed, the formulated goals and requirements for the pipeline, the proposed *visual analytics pipeline* (Section 3.2) comprising the data processing steps and the visual interface, and last the evaluation method. Following this, Section 4 presents the results of the models, the visual interface and the evaluation. Finally, *Discussion* (Section 5) and *Conclusions* (Section 6) are outlined.

¹ <https://liu.se/en/research/ai4climateadaptation> (last accessed: May 4, 2025)

2. Related work

2.1. Social media visual analytics

Social media visual analytics has gained significant attention in recent years, ranging from the more general tasks of assisting users in search and exploration of relevant information or understanding user behaviour [9] to various domain applications in social sciences, journalism, politics, urban planning, and crisis and emergency management [10]. Such approaches typically make use of the data associated with individual social media posts and their authors, including textual (and in some cases photo/video) contents, spatio-temporal attributes, and networks formed by reposts. Furthermore, additional data processing steps and computational analyses are commonplace, such as topic or sentiment analysis for text data. Visual analytics solutions have been developed, for instance, to monitor the key terms and sentiment in tweets [11] or clusters and representative phrases for several simultaneously occurring major events on Twitter [12].

The studies especially relevant to our project focus on using social media as a data source for supporting situational awareness for crisis and emergency management. The SensePlace2 system [13] supports spatio-temporal analysis of tweets collected via querying and processed with named entity extraction and geocoding; the interactive visual interface comprises multiple coordinated views such as query and history panels, a timeline, a map view, and a detailed tweet list. The workflows are initiated by either a query or interaction with one of the views. These design patterns are found across many spatio-temporal visual analytics tools relevant to our project. SensePlace3 [14] extends the suite of available analytical methods by including the extraction of implicit geolocation references and entity co-occurrence analysis. These solutions have been demonstrated and applied for multiple application scenarios. The Whisper interface by Cao et al. [15] complements this set of analyses with further techniques and visual representations for tracing information diffusion in real time, specifically for the scenario of an ongoing crisis.

2.2. Extreme weather event detection from VGI

While public authorities and organizations are the primary data sources for detection of extreme weather events and assessment of their impacts, alternative complementary sources have been explored. For example, the WeaVA approach presented by Haessig [16] supports interactive visual analysis of citizens' weather reports submitted to the Swiss national weather service. Similar to the more general approaches described above, WeaVA supports spatio-temporal data exploration with multiple coordinated views and faceted search, in this case facilitated by the structured data submitted by citizens. Meanwhile, this type of strategy has the limitation of being dependent on data collected via dedicated apps or portals.

Extracting relevant information to detect extreme weather events and their possible impacts from social media sources appears promising with respect to the volume of available data, especially assuming that users would be more inclined to share information on weather events perceived as extreme [17]. Extracting and classifying *relevant* posts and detecting the respective events are, however, challenging tasks. In recent years, data-driven AI methods have been applied for various stages of disaster management, including mitigation, risk of impact assessment, response, and recovery [18], some of which have relied on social media data. Rossi et al. [19] present an ML approach for early detection of impacts by using social media streams to identify extreme weather events and predict potential flood scenarios. Kankanamge et al. [20] highlight the application of an ML method in analysing social media data, specifically Twitter posts to identify and assess disaster severity as perceived by local communities. Hao and Wang [21] employed a machine learning approach that integrates the use of multimodal social media data to assess the risks of disaster impact. Rizk et al. [22] compare several multimodal pipelines for classifying tweet texts and images with respect to disaster impact assessment (nature vs infrastructure damage). Haung et al. [23] describe a neural network-based multimodal approach for classifying tweet texts and images as on/off-topic with respect to flood events.

Machine learning (ML) is essential for classification of VGI. In recent years the Bidirectional Encoder Representations from Transformers (BERT) architecture has been widely employed to process disaster-related social media text data [24]. Moreover, approaches combining text and image classification have also been proposed. Yan et al. [25] propose a framework that combines multimodal data, utilizing a BERT-BiLSTM-CRF model to identify urban flood locations and extract water depth information through text and image analysis. Brian-Smith et al. [26] utilize transformer networks and CLIP (Contrastive Language–Image Pretraining) to improve situational awareness during flood events. Although these approaches employ ML methods to perform the pursued tasks, they lack an advanced visual exploration interface to support analysts and other end-users in interactive investigation of the classified data.

The visual analytics research community has approached this challenge in several ways over the past years. Besides the solutions mentioned above, the study by Fuchs et al. [27] relies on spatio-temporal filtering and keyword filtering to identify flood related tweets which are then processed with a spatio-temporal clustering technique. By combining computational methods with map- and time-oriented visual representations, the analysts are enabled to explore, interpret, and validate findings in such disaster-related posts. The task of interactive event discovery in crisis-related tweets is also supported by ScatterBlogs [28] comprising spatio-temporal clustering, event-specific classifiers, and topic modelling. Interactive analysis of social media posts through the lenses of space, time, and textual contents (topics and keywords) was also approached by the study by Li et al. [29], who applied their semantics-space-time cube framework to a data set of flood-related tweets in English initially collected with keyword search. Snyder et al. [30] explored the integration of user-interactive ML models in combination with visual analytics tools. Their extension of the SMART toolkit allows users to iteratively retrain models to improve predictive accuracy [31].

Previous work by the authors on the topic includes initial experimentations with only text data from social media. Classic ML approaches for text classification were compared and a simple visual exploration interface was developed producing the first iteration of the pipeline [7]. The lessons learned from this initial work and further experimentation that included image classification led to the formulation of a design space of considerations and opportunities [8] for the subsequent creation of the VA pipeline presented in this paper.

3. Materials and methods

The work presented in this paper is the result of a research project concerned with exploring the potential of employing AI-based text and image analysis and visualization to support the process of impact-based weather warnings. The project was a collaboration between Linköping university, the Swedish Meteorological and Hydrological Institute (SMHI), and the county administrative board of Östergötland, Sweden. In 2021 the national system for impact-based weather warnings was introduced which requires a consultation process with local, regional, and national authorities before issuing certain types of weather warnings. The project has been running in parallel with this transition and has been continuously informed by feedback from project partners and other practitioners that were involved in the new weather warning system.

The project built on a co-design process to guide the formulation and development of the pipeline [6]. Continuous dialogues with the project partners (SMHI, County administrative board Östergötland) as part of workshops and project meetings have been a central driver in this design process for outlining design goals regarding appropriate functionality and corresponding analytical tasks. Moreover, the co-design process was further informed through a series of interviews and workshops with practitioners that were involved in the implementation of the impact-based weather warning system.

Twelve interviews were conducted in autumn/winter 2021–2022 with representatives of local, regional, and national authorities involved in the implementation of the new warning system. In the fall of 2022 a workshop was held with 22 participants as part of the national network meeting for regional climate adaptation coordinators. Another workshop was held in spring 2023 as part of the County administrative board's regional training program on impact-based weather warning systems, with 15 representatives from 9 different regional authorities [6]. The invaluable feedback gained through these events has been key for continuous refinement and validation of the visual analytics pipeline presented in Section 3.1. Finally, we conducted an evaluation session (Section 3.3) with 14 meteorologists and hydrologists with practical expertise in the weather warning system.

3.1. Analysis tasks and system requirements

Through the project's co-design process, we formulated a set of 4 overarching goals (G1–G4) for the analytic pipeline and from these extracted a set of 10 requirements (R1–R10), some referring to the machine learning functionality and some to the visual interface.

G1. Detection of extreme weather events from VGI texts and images. Using VGI, such as social media, can provide a new direct information input channel to the current weather warning process. Identifying the occurrence of extreme events and obtaining information regarding their severity can provide feedback regarding the validity and timeliness of the issued forecasts and warnings. The detection of extreme weather events from VGI posts requires the identification of a relevant subset of data through:

- R1.** *Text classification* for detection of VGI posts relevant to a selected type of extreme weather events (in our case flooding events).
- R2.** *Image classification* of VGI images to detect those relevant to a selected type of extreme weather event.

G2. Characterization of extreme weather events in search of local impacts. The relevant subset of VGI data identified through G1 needs to be further processed to extract relevant information regarding local impacts of weather events for further analysis. Such information includes the time, location, and further contextual characteristics of the weather events.

- R3.** Associate a timestamp with each VGI post to be able to explore periods of high post concentration as well as potential local temporal trends.
- R4.** Assign a geographic location to VGI posts to be able to explore the spatial extent and spatial variations of the weather events.
- R5.** Identify contextual characteristics from the VGI posts, texts and images, that can reveal potential impacts of the weather events.

G3. Multifaceted exploration of the spatial, temporal, and contextual characteristics of the detected extreme weather events. This goal refers to creating an appropriate visual interface allowing visualization and exploration of the VGI posts and their surrounding information. The main requirements relating to visualization as identified via discussions with project partners and stakeholder workshops/interviews are:

- R6.** Identification and exploration of the occurrence time and temporal extent of extreme weather events.
- R7.** Identification and exploration of the occurrence location and spatial extent of extreme weather events.
- R8.** Identification and exploration of the types of impacts of extreme weather events.
- R9.** Possibility to relate issued warnings to detected extreme weather events.



Fig. 2. Data processing pipeline integrated in the overall visual analytics pipeline developed in our project.

Table 1

Query of flood related Swedish keywords used for retrieving relevant social media posts and its translation.

"atmosfärisk flod" OR "hög vatten" OR åskskur OR regnskur OR dagvattensystem OR dränering OR "höga vågor" OR "höga flöden" OR dämmor OR snösmältning OR blött OR oväder OR stormflod OR vattenstånd OR vattennivå OR åskväder OR regnstorm OR "mycket regn" OR "kraftig regn" OR översvämningsskador OR översvämningar OR översvämning
"atmospheric river" OR "high water" OR thunderstorm OR rain shower OR "storm water system" OR drainage OR "high waves" OR "high flows" OR dams OR snow melt OR wet OR "bad weather" OR "storm surge" OR "water level" OR "water level" OR thunderstorm OR rainstorm OR "a lot of rain" OR "heavy rain" OR "flood damage" OR floods OR flood

G4. Human-in-the-loop in the machine learning pipeline. While the aim of the project has been to create a pipeline that offers automation capabilities for the identification of extreme weather events from VGI, such as social media, the input of the human expert to inform and/or validate the automation results is central. Interactive visualization of the results through G3 allows for such verification. An additional requirement aiming to improve the automation results is to:

R10. Allow human experts to provide feedback on the text and image classification assignments.

Based on these formulated goals and pipeline requirements, we implemented the proposed visual analytics pipeline in an iterative manner.

3.2. Visual analytics pipeline

Initial considerations regarding the design space and opportunities towards the creation of the proposed AI4CA VA pipeline were explored in previous work [7,8] where general directions and implementation possibilities for each step were deliberated.

The final VA pipeline integrates techniques from data science, natural language processing (NLP), image classification, and interactive visualization. The pipeline is designed to identify and visually explore (extreme) flooding events from VGI data, in particular social media data, with the greater aim to provide support to the SMHI impact-based weather warning system. In the following, the proposed VA pipeline, composed of the data processing pipeline (Fig. 2) and the visualization interface, for systematically analysing social media posts is described in more detail.

3.2.1. Data collection and pre-processing

The first step in the pipeline is concerned with the acquisition and pre-processing of VGI in the form of social media data posts for training and testing purposes.

A labelled data collection was retrieved from the CrisisLex repository [32], CrisisLexT6, which included tweets posted during 6 crisis events. The two flood related post collections from CrisisLexT6, the 2013 flood events in Alberta and Queensland, were extracted as separate datasets and used in this work for training and validation. These datasets contain approximately 16,000 records, featuring the tweet's unique ID, its textual content, and a label indicating the relevance of the tweet to the corresponding flood event as "on-topic" or "off-topic". An 80/20 split of this data was used for training and testing respectively. Further, for validating our pipeline, unlabelled social media posts were retrieved from the Twitter API v2 (via the then-available Academic Research track). A query of flood relevant terms was defined in a project workshop held together with experts from SMHI. Using this keyword query, social media data was collected for the period January 2015 until 2021 from Sweden resulting in 7046 unique posts (see Table 1).

The retrieved text data, from CrisisLex and the Swedish flooding posts, were pre-processed to prepare them for the text classification tasks.

In addition to the social media texts, several image datasets were also acquired for use in training and testing of the image classifiers. The publicly available European Flood 2013 Dataset [33] was used for a collection of "on-topic" labelled images. While for "off-topic" images, a selection of different images representing random scenes and objects was downloaded from the PASCAL

Visual Object Classes Homepage² [34]. The combined dataset (3646 images) was pre-processed by applying random resized crops to the images, converting them to tensors, and normalizing their pixel values. The data was divided into a 65/35 split for training and testing purposes respectively.

3.2.2. Text classification

Text classification is applied leveraging on NLP and ML to understand and categorize the semantic content of the social media post texts. The text classifier has been implemented by using a model based on BERT (Bidirectional Encoder Representations from Transformers) [35]. The variant used in this work is DistilBERT [36], the key benefit of this model lies in its ability to match BERT's performance while being significantly smaller and more efficient. The model was enhanced with transfer learning techniques using the training datasets retrieved from CrisisLex (as described in Section 3.2.1).

The labelled posts of the pre-processed training dataset were first tokenized using the DistilBERT tokenizer. Tokenization involves breaking down each post into individual words that can be fed into the language model. A custom data collator was then created for further preparing the input data for training the language model. The data collator was used to pad and truncate the tokenized posts so that they can be fed into the language model, since the model requires input sequences of fixed length. Following this, we trained the DistilBERT language model on the labelled training dataset to perform classification tasks to determine whether a tweet is “on-topic” or “off-topic”.

3.2.3. Geo-parsing

After setting the “on-topic” threshold to 0.5 and selecting all the post texts with a classification probability above 0.5 as flood-related, the posts were associated with a geographic location. Since most posts do not have an explicit geographic reference, geo-parsing was performed. Geo-parsing is concerned with using NLP techniques such as Named Entity Recognition (NER) to extract and assign a geographic location to the social media posts from the post texts. In this work, posts are processed through a geo-parsing script utilizing Mordecai v3 [37]. This geo-parsing step enables the geographical contextualization of social media posts, allowing for spatial analysis and mapping of the data.

3.2.4. Image classification

The social media posts associated with images were further processed by applying image classification to the corresponding images. While experimenting initially with convolutional neural network (CNN) based approaches, eventually a Vision Transformer (ViT) was used for the image classification step [38]. The reason for this is the overall performance of ViTs over CNNs with respect to computational resources [39] and their ability to accurately classify images of complex crowded environment such as cityscapes [40].

The pre-trained ViT model from Google, `vit-base-patch16-224-in21k`, was used in this work [38]. A training dataset with a 40/60 ratio of on-/off-topic images (887 on-topic, 1464 off-topic) was prepared using the collected image datasets (European Flood 2013, PASCAL VOC) as described in section *Data Collection and Pre-processing*. The model was trained on the pre-processed image data using the AdamW optimizer with a learning rate of 5×10^{-5} . The training process was run for three epochs. The logits (logarithmic probabilities) and true labels were then extracted from the prediction output. Finally, the `argmax` function was used to convert the logits to class predictions, and return both the predicted labels and the true labels as output.

3.2.5. Visualization interface

A visual interface was designed for exploring the output from the data classification pipeline; the social media posts reflecting flooding events. The interface is designed as a web-based application composed of 5 coordinated views showing different aspects (facets) of the data as shown in Fig. 3:

- (a) **Map View:** displays the spatial distribution of the posts.
- (b) **Image Cloud View:** shows the images associated with the posts, if any.
- (c) **Word Cloud View:** visualizes the most prominent terms in the explored posts.
- (d) **Posts View:** displays the text content of the posts.
- (e) **Timeline View:** shows the temporal distribution of the posts.

Map. In the map view (Fig. 3(a)) the collected posts are shown as blue dots on a 3D map. The blue dots are drawn transparent allowing for concentrations of dots to become visible as more opaque blue areas on the map. The map can be moved, rotated, and zoomed. The map acts as a filter to the explored social media post data. Posts visible on the map are the ones explored in the interface. Double clicking on a dot will select this post and the map will automatically zoom close to that dot, if the post includes an image, this image will be shown. A green rectangle indicates that the post is selected. The interaction mode changes when a post is selected; the post becomes locked in the centre and the map now rotates around that post. Double clicking on the map away from the post will reset the interaction back to the default. Clicking the home button resets the map to its original state.

The positions of the posts on the map carry a level of uncertainty since their locations are estimated from the text. To highlight this uncertainty, posts are not directly placed on the street level, but instead are placed 500 meters above the ground. Doing so, implies that when a user zooms in to the street level uncertain posts are no longer visible and cannot be perceived as exact locations of an event.

² <http://host.robots.ox.ac.uk/pascal/VOC> (last accessed: May 4, 2025)

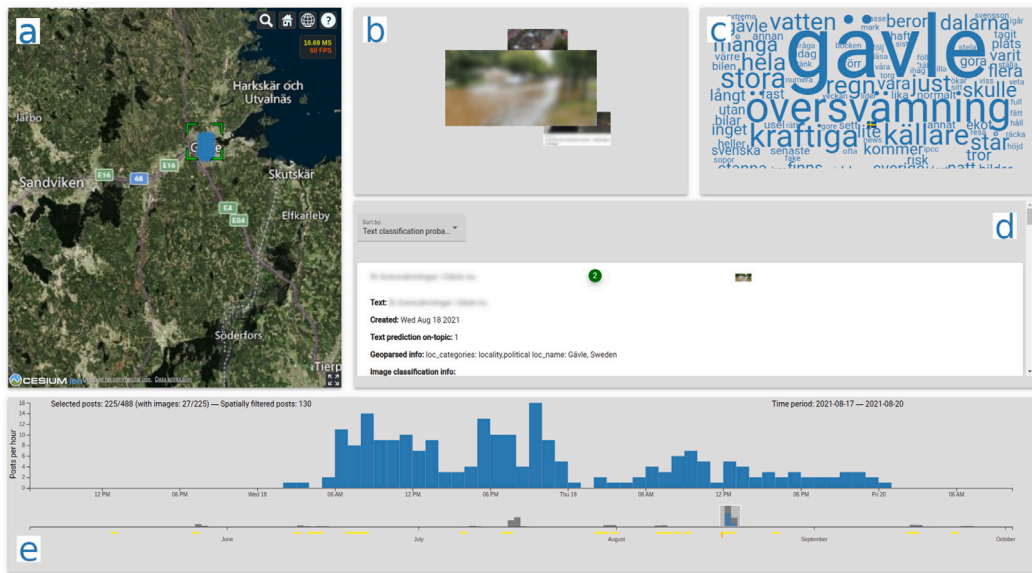


Fig. 3. Interactive visual interface composed of (a) the **map view** displaying the spatial distribution of the posts, (b) the **image cloud view** showing the images associated with the posts, (c) the **word cloud view** displaying the most prominent terms in the explored posts, (d) the **posts view** that displays the text content of the posts, and (e) the **timeline view** showing the temporal distribution of the posts. The lower histogram of (e) is the *overview timeline* showing the distribution of posts over the entire period May–September 2021 with the time range 17–20 August selected and highlighted, the upper histogram of (e) is the *filtered timeline* showing the temporal distribution of posts in the selected period in more detail. The linked views (a), (b), (c) and (d) visualize post facets from the filtered time range. The posts' images and text have been blurred to ensure that no restricted content is disclosed.

Image cloud view. The image cloud view (Fig. 3(b)) displays the images (if any) of the posts visible on the map. The view can be rotated, zoomed, and panned. It contains two distinct clusters: one contains images classified as flooding relevant and the other as flooding irrelevant. Clicking on an image in the view will select it. The image will be centred in the view and its position will be highlighted on the map by a green rectangle. When selected, the image becomes locked in the centre and the other images rotate as satellites around the selected image. Double clicking with the right button away from the image will reset the view and the interaction back to the default.

Word cloud view. The word cloud view (Fig. 3(c)) displays the prominent terms in the posts that are visible on the map. The size of the words reflects the frequency of a term; the more frequently a term is occurring, the larger the term will appear.

Posts view. The posts view (Fig. 3(d)) displays the posts visible on the map. The posts are displayed in a list with the first line of the post text visible. If the post has an image, this is visible as a thumbnail in the list. The posts are sorted by posting time by default. But can also be sorted by the classification probability of the texts or images respectively. When clicking on a post in the list, details about the post will be shown including the full post text, date as well as text and image classification probability. A classification feedback button is placed beside each post on the list. A user can leave feedback regarding the assigned text and image classification by clicking on the feedback button. A window pops up where an assessment of the classification can be entered.

Timeline. The timeline view (Fig. 3(e)) is composed of two histograms showing the distribution of posts over time. The *overview timeline* is placed below and shows an overview of the posts over the entire time period of the data. Here, each bar corresponds to a day and its height reflects the total number of posts per that day. The SMHI issued warnings for the time period are marked below the timeline axis with lines in the colour corresponding to the type of warning (yellow, orange, red). The *filtered timeline* is placed above, showing the posts for a user-selected time period. In this view each bar corresponds to an hour and its height reflects the total number of posts per that hour. A user can select a time period in the overview timeline using a time window filter and the data displayed on the upper filtered timeline will be updated accordingly. The overview timeline reflects the occurrence time of the posts visible on the map. When zooming into an area on the map, as seen in Fig. 3(a), the occurrence times of the filtered posts will be highlighted in the overview timeline by filling the day bars in blue (Fig. 4). This way relations between the spatial and temporal distributions can emerge.

3.3. Evaluation method

The pipeline was partly evaluated through continuous feedback collected on the iterations of the pipeline and interface as part of the co-design workshops of the project. In addition, a final systematic empirical evaluation of the visual interface was also conducted, aiming to: (a) assess the current exploration functionality in terms of content and interface features, and (b) identify any additional

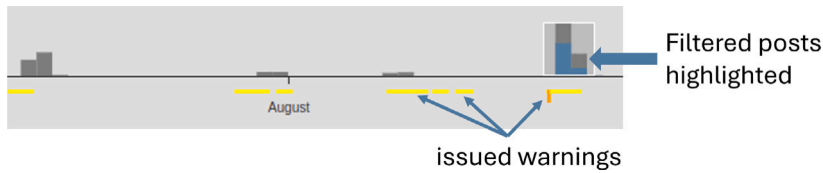


Fig. 4. Filtered posts highlighted in the overview timeline.

exploration refinements needed and formulate potential future directions. Structured user tests were conducted with meteorology and hydrology experts who performed tasks on the visual interface under the observation of research team members.

3.3.1. Setup and participants

14 experts from SMHI participated in the experiment, 10 meteorologists and 4 hydrologists, of which 8 were female and 6 males. The experiments took place at the SMHI premises. Two conference rooms were setup with a laptop connected to an external keyboard and mouse. We conducted 2 user test sessions in parallel over the course of two days. Each test session included an individual participant from SMHI and two research team members. Each test session lasted approximately 60 min. Each participant received an introductory text of the project and a video instruction of the visual interface ahead of the experiment. During the experiment they had access to a printed instruction handbook of the interface, in English. During the test sessions screen and audio recording was used.

3.3.2. Evaluation protocol

The experiment session started with a short introduction to the project and the test setup for the participant. Participants were asked to sign a consent form allowing the audio-recording of the test sessions. Following this, the video instruction of the visual interface was shown and the participant was asked whether they had any questions. A printed copy of the experiment protocol was handed out with instructions on the tasks to be performed. The participants were instructed to follow the task descriptions and perform the tasks using the interface. They were encouraged to think aloud while interacting with the interface and were reminded to do so during the experiment if needed.

Two sets of tasks were prepared for the experiment sessions. First, the participants performed tasks on each view of the interface separately (single view tasks) to successively test and get acquainted with the functionality of the tool. Following this, two combination tasks were performed that were aiming to allow the user to perform a full exploration using all views. After the tasks were completed, participants were asked to fill in a System Usability Scale (SUS) questionnaire [41] that was adjusted to refer to the visual interface in focus. This aimed to summarize participants' experiences and measure the perceived usability of the interface. As a final question, participants were asked "to what extent and how the proposed interface/pipeline would be relevant and usable in their work processes" to obtain their reflections on how the proposed pipeline would fit in their own work contexts.

4. Results

This section presents the results of the presented work, including the results of the model performance, a potential usage scenario exemplifying the proposed VA pipeline, and the results of the conducted interface evaluation.

4.1. Model results

The classifiers demonstrated good performance, achieving high accuracy, precision, and recall metrics; highlighting the effectiveness for the task. Both text and image classifiers were evaluated by calculating standard metrics [42]: accuracy (1), precision (2), recall (3), and F1 score (4).

Accuracy: The proportion of correctly identified instances out of the total instances.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \quad (1)$$

Precision: The proportion of true positive instances among all instances predicted as positive.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

Recall: The proportion of true positive instances among all actual positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

F1 Score: Harmonic average of precision and recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Table 2
Evaluation metrics for pipeline classifiers.

	Accuracy	Precision	Recall	F1 score
text classifier	0.973	0.952	0.980	0.966
image classifier	0.984	0.987	0.9825	0.985

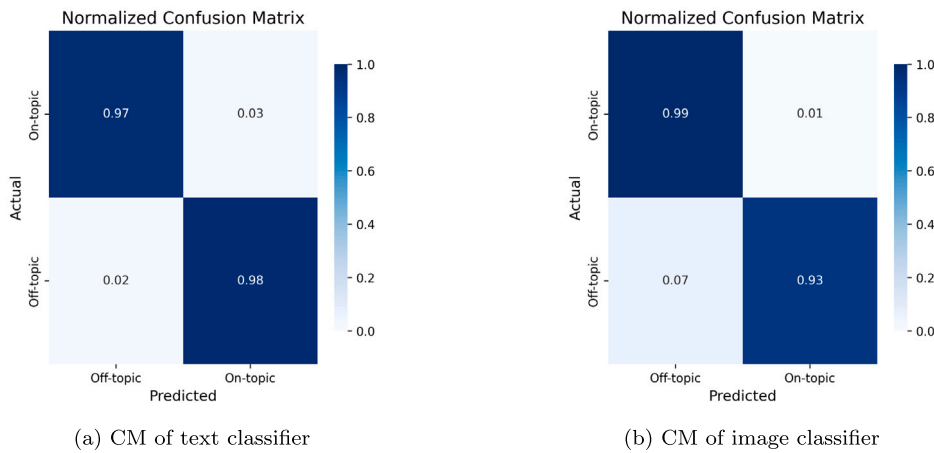


Fig. 5. Confusion matrix (CM) for pipeline (a) text and (b) image classifiers.

The performance metrics were derived via out-of-sample testing, where text and image classifiers were trained on a dataset and then evaluated on an unseen subset of that same dataset, thereby providing a more accurate assessment of their performance in real-world scenarios. The metrics results can be seen in Table 2, Figs. 5(a) and 5(b).

As can be seen from the results above, the models for both text and image classification tasks demonstrate high performance according to the standard classification metrics. However, the existing work shows that relying only on ML classification metrics may not be sufficient for actual deployment and application of such models for further downstream tasks. As Liao et al. [43] state, summarizing the model performance with a single number creates an illusion of simplicity that ignores many underlying assumptions in the learning problem. Hutchinson et al. [44] make a distinction between *learner-centric* evaluations (i.e., focusing on the ML algorithm/model) and *application-centric* evaluations (which are concerned with how the model will operate within an ecosystem comprising human actors/users and technical components). The study by Papenmeier et al. [45] with text classification tasks indicates that the *perceived* accuracy of ML models, as reported by human users, depends on whether the model makes mistakes with the cases that a human would find easy/difficult/impossible to classify. A very recent study conducted by Liu et al. [46] for image classification tasks with real-world images suggests that humans and ML models make significantly different mistakes, implying the potential for complementary teaming.

The need to provide capabilities to investigate the ML predictions in the wider context beyond benchmark performance/metrics was precisely the motivation for developing our visual analytic approach, while the results of combining both ML-based and interactive visual analyses are described next.

4.2. Usage scenario

The visual interface makes it possible to flexibly explore the outputs of our data processing pipeline. To illustrate how this can be done, we describe an example exploration scenario of a recent impactful extreme weather event in Sweden. The night between 17–18 of August 2021, Gävle municipality was affected by extreme rainfall. A total of 161.7 mm of rain fell within 24 h, with an intense 101 mm recorded in just two hours. This led to widespread flooding, which inundated low-lying areas, several roads became unusable or inaccessible, and approximately 4,000–5,000 homes were affected [47].

Processed posts for the period May to September 2021 are loaded into the visual interface. Two peaks are detected in the overview timeline (Fig. 3(e)), one in July and one in August. Choosing the higher peak period in August and inspecting the map clearly shows a higher concentration of posts in the Gävle area on the map. Initially, the word and image cloud display items corresponding to posts across Sweden from the selected period. A key interest in the exploration of this data is the local impact of the event, in particular regarding the identification of potential impacts to urban infrastructure. Zooming and moving the map updates the image and word cloud views and focuses the exploration enabling the closer inspection of such impacts. The images show evidence of the impact on roads and urban infrastructure, while the word cloud provides additional information concerning flooded basements, blocked road drainage, and cars being stuck. Further details can be accessed by visiting the posts to obtain the full post texts, the geoparsing results as well as the classification results.

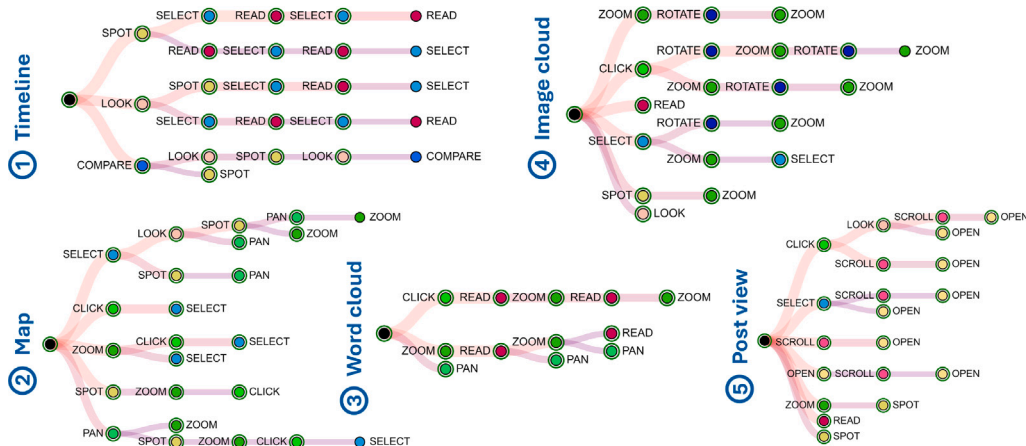


Fig. 6. Interaction patterns identified for each of the individual views of the interface: (1) timeline view, (2) map view, (3) word cloud view, (4) image cloud view, (5) post view. The nodes are categorically coloured to distinguish the different keywords.

4.3. Interface evaluation results

The material collected through the user evaluation (Section 3.3) was analysed to inform the relevance, usability and usefulness of the visual interface.

4.3.1. Interaction pattern analysis

We processed the test session recordings (video and audio) of the participants and extracted interaction logs from them to quantify and summarize their main interaction patterns.

The recordings were initially reviewed by one of the research team members and a set of characterizing interaction keywords were extracted to capture the participants' interactions with the visual interface. The keywords were CLICK, COMPARE, COUNT, LOCATE, LOOK, OPEN, PAN, READ, ROTATE, SCROLL, SELECT, SPOT, ZOOM and they were retrieved by examining the interactions of the participants in the video as well as listening to their explanations of what they were doing from the think aloud audio track. The keywords were chosen to describe the actions at quite a low level and could therefore have adjusted meanings for the different views indicating the same interaction action being used in different contexts, e.g. SELECT a time range in the timeline view, or SELECT an image in the image cloud view. A participant's actions could then be reduced to a set of interaction sequences for each task composed of these keywords only. For example, the action sequence "SELECT the time range June–August on the lower timeline. LOOK at the concentration of posts on the upper timeline and SPOT a peak, results in the interaction sequence: SELECT → LOOK → SPOT.

A coding instruction was prepared outlining the transcribing process which was used by four of the authors to translate all 14 participant test sessions into interaction sequences. Only the set of single view tasks from the evaluation protocol was coded. This was done so that interaction patterns for each of the views could be explored separately. For each participant, the interaction sequences of all tasks per individual view were merged to one sequence. This process resulted in five datasets of interaction sequences, one per view of the interface, each of which was composed of one interaction sequence per participant.

We used the interactive sequence mining approach proposed by Vrotsou and Nordman [48] to explore and identify common interaction patterns corresponding to each of the views of the interface. The approach is using an adjusted pattern-growth mining algorithm which allows users to control the mining process by setting local constraints and growing patterns interactively in a pattern tree representation. We explored patterns with a 50% support allowing at maximum 1 interrupting activity between pattern activities (gap constraint). This process helps us to summarize the common interaction approaches used for each view by the participants as well as compare whether these correspond to the interactions anticipated in the interface design. Worth mentioning is the fact that the proposed interaction pattern analysis can be easily generalized and applied to any other visual exploration tools, since it is platform independent and based on a screen recording annotation process.

The resulting interaction patterns for each view under these constraints can be seen in Fig. 6. The patterns are displayed as a pattern tree, the branches are sorted according to the pattern support which is also represented on the thickness of the links. The interaction with most views matched the expected behaviour and participants did not have difficulties navigating the interface. For example, the most common pattern on the timeline view was for participants to spot the peaks of posts on the timeline, select a time range corresponding to each peak and read the corresponding frequency on the axis of the plot (SPOT → SELECT → READ → ...). Of more interest were the interactions patterns exhibited in the image cloud view, which was the view that was least familiar to the participants. Here, the expected interaction was to navigate the view through rotating and zooming in/out to explore the content of the images, and then select an image to explore the content of the corresponding post in the post view. While this was indeed the most common mode of interaction, participants developed an alternative interaction strategy. They would select an image from

one of the image clusters, which would bring it to focus and change the interaction mode so the surrounding images are rotating as satellites around it, zoom slightly out and then rotate to bring more of them in view. This pattern was also noticed by the observing researchers during the experiments as an unexpected behaviour but an effective interaction alternative for the particular view.

4.3.2. System usability scale responses

The SUS questionnaires filled in by participants were analysed using the System Usability Scale Analysis Toolkit [41]. The analysis showed a SUS study score of 77.69 (*median* = 82.5, *SD* = 10.67) which indicates a “Good” usability according to the adjective contextualization scale [49] or a “Grade B” according to the letter grade scale and is an *acceptable* usability score. The sample size of $n = 14$ participants yields results that are expected to be 100% conclusive [50].

4.3.3. Relevance and usability

After the completion of the entire test session, participants were asked to reflect about the relevance and usability of the pipeline in their work processes.

In response to this question, all but one participant raised *validation of issued weather warnings*. Participants argued that the data provided through the pipeline would supply valuable information on observed impacts of previous events and could hence support the validation of the geographic extent and intensity of a weather event to which a warning had been issued. One respondent raised that validation of issued warnings is part of their work tasks, which involves to research media and sometimes also social media data, and that the pipeline would make this process easier. In the post-event evaluation, data on impacts was considered to be hard to allocate, and social media-based data on impacts would provide better insight into what happened in the real location. The information for validation could thus be an addition to the feedback from the county administrative board, which, as one respondent noted might not always be provided nor sufficiently informative. Two respondents raised the issue that response and observation data is usually scarce, and that this additional data source could improve the assessment of what happened on the ground, and further support future analysis of impacts and risks of certain weather events and thresholds.

Six of the respondents raised specifically the *relevance of real-time data during an issued weather event*, while some of the respondents emphasized that real-time data provided by social media would not influence the decision whether a warning would be issued. They argued that it nevertheless would be relevant to see any possible impacts, which cannot be detected through earth observations (satellites/radar), and in this regard, the pipeline would provide a relevant source of additional information for, what one respondent referred to as ‘nowcasting’. Several participants specifically raised the value of having additional information to the modelled data, in particular when referring to ‘hard-to-predict’ events, such as heavy rainfall, where the intensity in specific geographic locations is uncertain. In such cases, one of the respondents argued, social media data, but also news reports might provide important information about both impacts and more precise locations. Another respondent also raised the relevance for confirmation whether an event is still ongoing, in which case a warning should not yet be removed.

While some respondents raised the point that the system would nevertheless require an assessment of the relevance and reliability of reports provided by private citizens, the general features of the proposed pipeline were considered as useful for both validation and support during ongoing events, as the *larger coverage that crowd-sourced information could provide* was considered as relevant complementary data to the weather stations and other available observations. Finally, the *provision of a more detailed time span of the warnings*, and the *geographic extent of the events and their impacts*, allowing the user to connect issued warnings with impacts in one single interface, was raised as valuable by several respondents.

5. Discussion

The proposed VA pipeline is the result of an interdisciplinary collaboration between researchers and practitioners aiming to address real-world processes relating to prediction and impact assessment of extreme weather events through the use of new technologies and non-standard information sources. Central to this interdisciplinary collaboration has been the regular contact and continuous dialogue between researchers and practitioners to ensure that a shared focus and common vocabulary were formulated and maintained. This provided a better understanding of the current workflows and the opportunities for improvements and guided the scope, requirements and direction of the research process.

Practitioners acknowledged the pipeline itself as a valued contribution to impact-based warnings, and regarded the proposed system as a highly useful tool that could provide them with functionality suitable for validating weather warnings. As such, this could with advantage be added to the suite of tools they are currently using as part of the impact-based warning processes. SMHI practitioners reported that they currently use news and social media outlets as input both during an issued warning and in the post-evaluation process to identify potential local impacts on infrastructure and population and to better understand vulnerability and exposure. Various news and social media are manually scanned for relevant posts which are then individually classified. This manual process is hard to achieve on a large scale, and results are gathered as small datasets with a low level of detail. The proposed pipeline provides this information from social media at an already classified level, allowing the domain expert to focus on their expert tasks in line with Shneiderman’s recommendations on Human-Centered AI [51].

There are, however, also certain limitations and issues relating to the proposed approach that need to be discussed. First, working with VGI collected from social media as a data source unavoidably involves concerns regarding the validity and representativeness of the collected datasets [52]. The accuracy and quality of the content can, for example, have a great influence on the detection of relevant events. Fake posts containing flooding related terminology or outdated images may bias the data and result in pseudo-events being identified by the proposed approach. This risk has been significantly amplified by the rise of generative AI, since the ability

to generate highly realistic images, videos, and text with ease has made it difficult to distinguish synthetic content from authentic reporting. Strategies and functionality for detecting and filtering out misleading content and unreliable sources should therefore be considered. Boididou et al. [52] have proposed a system that detects and filters out misleading content on Twitter by leveraging credibility-oriented features and a two-step classification model with semi-supervised learning. More recently, Wang et al. [53] proposed a framework to detect fake news by combining text, visuals, and a large language model, employing feature extraction, relationship refinement, and explanation generation to identify and filter misleading content.

Bias can also emerge in our proposed approach from the keywords used for identifying which posts are relevant to include in the analysis. Different users may use different terminology for describing similar weather events and impacts, while certain terms may be ambiguous, resulting in noise in the data. A careful selection of relevant keywords is crucial to get results that are relevant and representative of the event. In our work, we have relied on well established collections of relevant terms [32] which we have adjusted and complemented in workshops with practitioners.

As events of interest in our approach are identified through the spatio-temporal concentration of the collected VGI, the spatial distribution and density of the posts are another potential source of bias in the datasets. Naturally, locations with larger population density, such as urban areas, will also have larger representation in the data since more individuals may be creating posts regarding weather event impacts. While this confirms the relevance of exploring VGI for urban risk management, it also highlights a risk that less significant events with limited impact can potentially be exaggerated and emerge as more prominent due to a high concentration of posts whereas truly broad impact events in less populated areas may remain unnoticed due to a small number of posts. To amend this, approaches for assigning weighting factors to posts depending, for example, on impact severity detected from the post texts or source reliability could be of interest.

Most importantly, the accuracy of the geographic positions of the collected VGI is an issue that is crucial yet difficult to handle, both in our approach and in general. As most posts do not have an explicit geographic tag marking an exact location, NER-based approaches for inferring a location from the post texts are commonly used. The level of detail of the identified locations from such techniques is commonly low and the uncertainty of the predictions high. In our experience when testing and discussing the pipeline with practitioners working with impact-based weather warnings, this lack of location precision has been problematic as the tendency has been to consider it exact. Frequently, experts zoomed in on the post locations on the map trying to identify the affected infrastructure mentioned in the post or included in the corresponding image. While strategies could be implemented for increasing the quality of the predictions, the approach will never provide the expected accuracy and reliability. It is, therefore, crucial to make sure that this uncertainty of the location information is reflected in the visualization to avoid misinterpretation. In the proposed interface we have experimented with vertically positioning the post markers on the map relative to the level of detail of the predicted location, at 500 m above the ground level as previously discussed.

Another issue that arises from the nature of social media as a data source in general is the uncertainty in their accessibility/availability over time. The APIs used for accessing the data may be updated from the provider resulting in interruptions in the data collection. Moreover, permissions may change rendering the data source inaccessible for research. While there is no permanent solution for this limitation, it is important to continuously stay up-to-date with the API communication and it could be beneficial to be exploring multiple media and social media as input sources.

Despite the potential that VGI from social media provide for the identification of extreme weather events, several of the limitations discussed arise from the fact that the data are used for a purpose they were not intended for. These limitations could be mitigated by introducing a dedicated VGI collection system for impact-based weather warnings. An interesting prospect would be to include, for example, functionality in existing weather apps of the national meteorological institutes that would allow citizens to contribute direct observations of weather event impacts in the form of images and text descriptions. This would potentially provide higher location accuracy and less noisy data. The proposed pipeline has also significant practical implications for urban environments. By leveraging on the social media data analysis and exploration, cities can better identify risk areas to be integrated in the assessment processes with the national system for impact-based weather warnings.

6. Conclusions

In this work, we have formulated, designed and tested a VA pipeline for the detection and visual exploration of extreme weather events and their impacts from VGI text and image data. The pipeline has been developed with a focus on flooding events but could be used on data concerning other extreme weather events such as heatwaves or severe thunderstorms. The VA pipeline can support local risk assessments by classifying and analysing VGI data from social media, providing insights that can help to better understand local vulnerabilities and exposure and assess areas at risk. It can also facilitate the communication of strategic climate adaptation measures, providing a tool for experts to share and discuss actionable strategies with stakeholders. Furthermore, while the pipeline has been tested with VGI from social media, data from weather or crowdsourcing apps could also be explored and analysed. Overall, the conducted evaluation indicated a good perceived relevance and usability of the visual interface and demonstrated the potential of the visual analytics pipeline as a valuable tool to be used as a complement in the current impact-based forecasting workflows of governmental agencies responsible for impact-based weather warning systems. Future work of interest includes optimizations of the visual interface for handling extensive number of events, for example by applying anti-cluttering techniques, such as dynamic glyph representation. Another relevant extension would be to include a component to identify extreme events from video data using multimodal large language models in the classification pipeline. Finally, a valuable addition to investigate could be to integrate historical weather data to help identify correlations between reports and weather conditions.

CRediT authorship contribution statement

Carlo Navarra: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kostiantyn Kucher:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Tina-Simone Neset:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Clara Greve Villaro:** Writing – review & editing, Resources. **Fredrik Schück:** Writing – review & editing, Resources. **Jonas Unger:** Writing – review & editing, Investigation, Funding acquisition, Conceptualization. **Katerina Vrotsou:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Funding

This research was funded by Sweden's Innovation Agency, VINNOVA, grant number 2020-03388, 'AI for Climate Adaptation'.

Declaration of competing interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgement

We want to thank the AI4CA project collaborators and the participants in workshops and interviews.

Data availability

The authors do not have permission to share data.

References

- [1] IPCC, Summary for policymakers, in: *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 2022, http://dx.doi.org/10.1007/978-3-662-66497-1_2.
- [2] R.R. Biswas, A. Rahman, Adaptation to climate change: A study on regional climate change adaptation policy and practice framework, *J. Env. Manag.* 336 (2023) 117666, <http://dx.doi.org/10.1016/j.jenvman.2023.117666>.
- [3] World Meteorological Organization, *WMO Guidelines on Multi-Hazard Impact-Based Forecast and Warning Services. Part II: Putting Multi-Hazard IBFWS Into Practice*, Report WMO-No. 1150, Part II, WMO, 2022.
- [4] World Meteorological Organization, *WMO Guidelines on Multi-Hazard Impact-Based Forecast and Warning Services*, Report WMO-No. 1150, WMO, 2015.
- [5] M.F. Goodchild, Citizens as sensors: The world of volunteered geography, *GeoJournal* 69 (2007) 211–221, <http://dx.doi.org/10.1007/s10708-007-9111-y>.
- [6] T.-S. Neset, K. Vrotsou, L. Andersson, C. Navarra, F. Schück, M.M. Edström, C. Rydholm, C.G. Villaro, K. Kucher, B.-O. Linnér, Artificial intelligence in support of weather warnings and climate adaptation, *Clim. Risk Manag.* 46 (2024) 100673, <http://dx.doi.org/10.1016/j.crm.2024.100673>.
- [7] L. Styve, C. Navarra, J.M. Petersen, T.-S. Neset, K. Vrotsou, A visual analytics pipeline for the identification and exploration of extreme weather events from social media data, *Climate* 10 (2022) 174, <http://dx.doi.org/10.3390/cli10110174>.
- [8] K. Vrotsou, C. Navarra, K. Kucher, I. Fedorov, F. Schück, J. Unger, T.-S. Neset, Towards a volunteered geographic information-facilitated visual analytics pipeline to improve impact-based weather warning systems, *Atmosphere* 14 (2023) <http://dx.doi.org/10.3390/atmos14071141>.
- [9] Y. Wu, N. Cao, D. Gotz, Y.-P. Tan, D.A. Keim, A survey on visual analytics of social media data, *IEEE Trans. Multimed.* 18 (2016) 2135–2148, <http://dx.doi.org/10.1109/tmm.2016.2614220>.
- [10] S. Chen, L. Lin, X. Yuan, Social media visual analytics, *Comput. Graph. Forum* 36 (2017) 563–587, <http://dx.doi.org/10.1111/cgf.13211>.
- [11] N. Diakopoulos, Diamonds in the rough: Social media visual analytics for journalistic inquiry, in: *Proc. of VAST*, IEEE, 2010, pp. 115–122, <http://dx.doi.org/10.1109/vast.2010.5652922>.
- [12] J. Knittel, S. Koch, T. Tang, W. Chen, Y. Wu, S. Liu, T. Ertl, Real-time visual analysis of high-volume social media posts, *IEEE Trans. Vis. Comput. Graphics* 28 (2021) 879–889, <http://dx.doi.org/10.1109/tvcg.2021.3114800>.
- [13] A.M. MacEachren, A. Jaiswal, A.C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, J. Blanford, SensePlace2: Geotwitter analytics support for situational awareness, in: *Proc. of VAST*, IEEE, 2011, pp. 181–190, <http://dx.doi.org/10.1109/vast.2011.6102456>.
- [14] S. Pezanowski, A.M. MacEachren, A. Savelyev, A.C. Robinson, SensePlace3: A geovisual framework to analyze place-time-attribute information in social media, *Cartogr. Geogr. Inf. Sci.* vol. 45 (2018) 420–437, <http://dx.doi.org/10.1080/15230406.2017.1370391>.
- [15] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, H. Qu, Whisper: Tracing the spatiotemporal process of information diffusion in real time, *IEEE Trans. Vis. Comput. Graphics* 18 (2012) 2649–2658, <http://dx.doi.org/10.1109/tvcg.2012.291>.
- [16] D. Haessig, H. Yang, R. Pajarola, A. Diehl, WeaVA: Weather event characterization based on citizen reports, in: *Proc. of EnvirVis, EuroVis, Eurographics*, 2024, <http://dx.doi.org/10.2312/envirvis.20241136>.
- [17] Y. Feng, X. Huang, M. Sester, Extraction and analysis of natural disaster-related VGI from social media: Review, opportunities and challenges, *Int. J. Geogr. Inf. Sci.* 36 (2022) 1275–1316, <http://dx.doi.org/10.1080/13658816.2022.2048835>.
- [18] W. Sun, P. Bocchini, B.D. Davison, Applications of artificial intelligence for disaster management, *Nat. Hazards* 103 (2020) 2631–2689, <http://dx.doi.org/10.1007/s11069-020-04124-3>.
- [19] C. Rossi, F. Acerbo, K. Ylinen, I. Juga, P. Nurmi, A. Bosca, F. Tarasconi, M. Cristoforetti, A. Alikadic, Early detection and information extraction for weather-induced floods using social media streams, *Int. J. Disaster Risk Reduct.* 30 (2018) 145–157, <http://dx.doi.org/10.1016/j.ijdr.2018.03.002>.
- [20] N. Kankaname, T. Yigitcanlar, A. Goonetilleke, Md. Kamruzzaman, Determining disaster severity through social media analysis: Testing the methodology with South East Queensland flood tweets, *Int. J. Disaster Risk Reduct.* 42 (2020) 101360, <http://dx.doi.org/10.1016/j.ijdr.2019.101360>.
- [21] H. Hao, Y. Wang, Leveraging multimodal social media data for rapid disaster damage assessment, *Int. J. Disaster Risk Reduct.* 51 (2020) 101760, <http://dx.doi.org/10.1016/j.ijdr.2020.101760>.

- [22] Y. Rizk, H.S. Jomaa, M. Awad, C. Castillo, A computationally efficient multi-modal classification approach of disaster-related Twitter images, in: Proc. of ACM SIGAPP, ACM, 2019, pp. 2050–2059, <http://dx.doi.org/10.1145/3297280.3297481>.
- [23] X. Huang, Z. Li, C. Wang, H. Ning, Identifying disaster related social media for rapid response: A visual-textual fused CNN architecture, Int. J. Digit. Earth 13 (2020) 1017–1039, <http://dx.doi.org/10.1080/17538947.2019.1633425>.
- [24] R. Prasad, A.U. Udemé, S. Misra, H. Bisallah, Identification and classification of transportation disaster tweets using improved bidirectional encoder representations from transformers, Int. J. Inf. Manag. Data Insights 3 (2023) 100154, <http://dx.doi.org/10.1016/j.jjime.2023.100154>.
- [25] Z. Yan, X. Guo, Z. Zhao, L. Tang, Achieving fine-grained urban flood perception and spatio-temporal evolution analysis based on social media, Sustain. Cities Soc. 101 (2024) 105077, <http://dx.doi.org/10.1016/j.scs.2023.105077>.
- [26] L. Bryan-Smith, J. Godsall, F. George, K. Ego, N. Dethlefs, D. Parsons, Real-time social media sentiment analysis for rapid impact assessment of floods, Comput. Geosci. 178 (2023) 105405, <http://dx.doi.org/10.1016/j.cageo.2023.105405>.
- [27] G. Fuchs, N. Andrienko, G. Andrienko, S. Bothe, H. Stange, Tracing the German centennial flood in the stream of tweets: First lessons learned, in: Proc. of GEOCROWD at SIGSPATIAL, ACM, 2013, pp. 31–38, <http://dx.doi.org/10.1145/2534732.2534741>.
- [28] D. Thom, R. Krüger, T. Ertl, Can Twitter save lives? A broad-scale study on visual social media analytics for public safety, IEEE Trans. Vis. Comput. Graphics 22 (2016) 1816–1829, <http://dx.doi.org/10.1109/tvcg.2015.2511733>.
- [29] J. Li, S. Chen, W. Chen, G. Andrienko, N. Andrienko, Semantics-space-time cube: A conceptual framework for systematic analysis of texts in space and time, IEEE Trans. Vis. Comput. Graphics 26 (2018) 1789–1806, <http://dx.doi.org/10.1109/tvcg.2018.2882449>.
- [30] L.S. Snyder, Y.-S. Lin, M. Karimzadeh, D. Goldwasser, D.S. Ebert, Interactive learning for identifying relevant tweets to support real-time situational awareness, IEEE Trans. Vis. Comput. Graphics 26 (2019) 558–568, <http://dx.doi.org/10.1109/tvcg.2019.2934614>.
- [31] J. Zhang, J. Chae, C. Surakitbanharn, D.S. Ebert, SMART: Social media analytics and reporting toolkit, in: Proc. of ViP at IEEE VIS, 2017.
- [32] A. Olteanu, C. Castillo, F. Diaz, S. Vieweg, CrisisLex: A lexicon for collecting and filtering microblogged communications in crises, in: Proc. of ICWSM, volume 8, AAAI, 2014, pp. 376–385, <http://dx.doi.org/10.1609/icws.v8i1.14538>.
- [33] B. Barz, K. Schröter, M. Münch, B. Yang, A. Unger, D. Dransch, J. Denzler, Enhancing flood impact analysis using interactive retrieval of social media images, Arch. Data Sci. A 5 (2018) A06, <http://dx.doi.org/10.5445/KSP/1000087327/06>, 21 S. online.
- [34] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2010) 303–338, <http://dx.doi.org/10.1007/s11263-009-0275-4>.
- [35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. of NAACL, ACL, 2019, pp. 4171–4186, <http://dx.doi.org/10.18653/v1/N19-1423>.
- [36] V. Sanh, L. Debut, J. Chaumond, T. and Wolf, DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, in: Proc. of EMC2 at NeurIPS, 2019, <http://dx.doi.org/10.48550/arXiv.1910.01108>.
- [37] A. Halterman, Mordecai 3: A neural geoparser and event geocoder, 2023, <http://dx.doi.org/10.21105/joss.00091>, arXiv preprint [arXiv:2303.13675](https://arxiv.org/abs/2303.13675).
- [38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16 × 16 words: Transformers for image recognition at scale, in: Proc. of ICLR, 2021, <http://dx.doi.org/10.48550/arXiv.2010.11929>.
- [39] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks? Adv. Neural Inf. Process. Syst. 34 (2021) 12116–12128, <http://dx.doi.org/10.48550/arXiv.2108.08810>.
- [40] A. Abdelraouf, M. Abdel-Aty, Y. Wu, Using vision transformers for spatial-context-aware rain and road surface condition detection on freeways, IEEE Trans. Intell. Transp. Syst. 23 (2022) 18546–18556, <http://dx.doi.org/10.1109/tits.2022.3150715>.
- [41] J. Blattgerste, J. Behrends, T. Pfeiffer, A web-based analysis toolkit for the system usability scale, in: Proc. of PETRA, ACM, 2022, pp. 237–246, <http://dx.doi.org/10.1145/3529190.3529216>.
- [42] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (2009) 427–437, <http://dx.doi.org/10.1016/j.ipm.2009.03.002>.
- [43] T. Liao, R. Taori, D. Raji, L. Schmidt, Are we learning yet? A meta review of evaluation failures across machine learning, in: Proc. of NeurIPS Datasets and Benchmarks, vol. 1, 2021.
- [44] B. Hutchinson, N. Rostamzadeh, C. Greer, K. Heller, V. Prabhakaran, Evaluation gaps in machine learning practice, in: Proc. of ACM FAccT, ACM, 2022, pp. 1859–1876, <http://dx.doi.org/10.1145/3531146.3533233>.
- [45] A. Papenmeier, D. Kern, D. Hienert, Y. Kammerer, C. Seifert, How accurate does it feel? — Human perception of different types of classification mistakes, in: Proc. of ACM CHI, ACM, 2022, <http://dx.doi.org/10.1145/3491102.3501915>.
- [46] M. Liu, J. Wei, Y. Liu, J. Davis, Human and AI perceptual differences in image classification errors, Proc. AAAI Conf. Artif. Intell. 39 (2025) 14318–14326, <http://dx.doi.org/10.1609/aaai.v39i13.33568>.
- [47] E. Glaas, M. Hjerpe, S. Storbjörk, The 2021 extreme rainfall in Gävle, Sweden: Impacts on municipal welfare services and actions towards more resilient premises and operations, Hydrol. Res. 55 (2024) 431–443, <http://dx.doi.org/10.2166/nh.2024.107>.
- [48] K. Vrotsou, A. Nordman, Exploratory visual sequence mining based on pattern-growth, IEEE Trans. Vis. Comput. Graphics 25 (2019) 2597–2610, <http://dx.doi.org/10.1109/tvcg.2018.2848247>.
- [49] A. Bangor, P. Kortum, J. Miller, Determining what individual SUS scores mean: Adding an adjective rating scale, J. Usability Stud. 4 (2009) 114–123.
- [50] T.S. Tullis, J.N. Stetson, A comparison of questionnaires for assessing website usability, in: Proc. of UPA, 2004, <http://dx.doi.org/10.1093/iwc/iwt019>.
- [51] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, Int. J. Hum. Comput. 36 (2020) 495–504, <http://dx.doi.org/10.1080/10447318.2020.1741118>.
- [52] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, Y. Kompatsiaris, Detection and visualization of misleading content on Twitter, Int. J. Multimed. Inf. Retr. 7 (2018) 71–86, <http://dx.doi.org/10.1007/s13735-017-0143-x>.
- [53] J. Wang, Z. Zhu, C. Liu, R. Li, X. Wu, LLM-enhanced multimodal detection of fake news, PloS One 19 (2024) <http://dx.doi.org/10.1371/journal.pone.0312240>.