Social Media Analysis for Situational Awareness and Impact Assessment in Flood Risk Management

Valerio Lorini

TESI DOCTORAL UPF / year 2022

THESIS SUPERVISOR
Prof. Dr. Carlos Castillo
Department of Information and Communications Technologies

Dedicated to people suffering from disasters

La lutte elle-même vers les sommets suffit à remplir un cœur d'homme;
il faut imaginer Sisyphe heureux
(A.Camus)

Acknowledgments

I started this Ph.D. journey some twenty years after I graduated in Computer Science with a thesis on Artificial Intelligence for planning. I want to thank those who made this possible, and I find it easier to do it in chronological order.
I was already working as a researcher, although on a different topic, when thanks to Milan and our first exploratory activity he designed, I had the opportunity to feel excited about crisis informatics. I got back where I felt at home, happy to go to the office each day. In the same project, I met Carlos, who accepted the challenge of supervising an 'older 'working' student. Little did I know how lucky I was, and perhaps little he knew how much I needed his help. There is a Japanese term that I'd like to use to describe his mentorship to me: 'Shibui', which refers to a particular aesthetic of simple, subtle, and unobtrusive beauty. That's my feelings every time I left our meetings after starting from an undefined mess of ideas. This defines him! a zen master ... or a Jedi, for the nerds. I want to thank Peter for letting me bring my studies into my work and my work into my studies, trusting and listening to me. I thank Marie and Noah for supporting me, letting me quietly work on weekends, and sacrificing holidays and time. I will always be grateful. I am also fortunate to have one of my parents read that I love them. I thank them for showing me that I could do what I wanted through dedication, hard work, and a little bit of awkwardness and craziness. I now know what I want. This Ph.D. reached beyond the mere research extent. Finally, I want to thank all the friends, colleagues, UPF staff that helped me during these years by just chatting, laughing, listening, and comforting me. Brightening my days.

Abstract

The interactions among people on social media are a form of distributed intelligence, as they allow people to make sense of a developing event collectively. Social media users can contribute to creating a 'sensor' for user-generated data that modeling or monitoring systems can assimilate during a crisis. However, social media platforms may not provide the functionality of summarizing useful information for crisis responders. We developed a platform to streamline the processing of text and images extracted from Twitter in near real-time during floods to solve this problem. Social media analysis can improve situational awareness in the form of a map or a report. When combined with risk analysis and socio-economic data, it could shorten the time needed to fill the time gap between the definition of the risk and the actual impact of a flood. Emergency managers could aggregate annotated data to confirm a forecast or monitor an event's development. Crisis responders can filter social media messages to distill specific needs when they must act quickly. Finally, we explore a quantitative integration of social media information into geospatial information systems to compute the flood extent in urban areas.

Resumen

Las interacciones entre las personas a través de las redes sociales son una forma de inteligencia distribuida, ya que permiten dar sentido a un evento en desarrollo de manera colectiva. Los usuarios de redes sociales pueden contribuir a crear un "sensor"para que datos generados por los ciudadanos puedan ser asimilados durante una crisis por sistemas de modelado o monitoreo. Sin embargo, las plataformas de redes sociales no ofrecen una funcionalidad para resumir la información útil para la gestión de una crisis. Para resolver este problema, desarrollamos una plataforma para agilizar el procesamiento de texto e imágenes extraídos de Twitter en tiempo casi-real durante inundacione. El análisis de redes sociales puede mejorar la conciencia situacional a través de mapas o informes. Cuando se combina con el análisis de riesgos y datos socioeconómicos, podría acortar el tiempo entre la definición del riesgo y el impacto real de una inundación. Los administradores de emergencias podrían utilizar datos anotados automáticamente para confirmar un pronóstico o monitorear el desarrollo de un evento. Las personas encargadas de la gestión de una crisis pueden filtrar los mensajes de redes sociales para destilar aquellos que atienden necesidades específicas, en particularmente en los casos en que deben actuar rápidamente. Finalmente, exploramos una integración cuantitativa de la información de las redes sociales en sistemas de información geoespacial para calcular la extensión las inundaciones en áreas urbanas.

Contents

List of Figures

List of Tables

Acronyms

CEMS Copernicus Emergency Management System. 7, 8, 17, 85–87, 90, 94, 98, 102, 103, 106

DEM Digital Elevation Model. 87, 89, 91, 92, 96–102, 128

DFO Dartmouth Flood Observatory. 26–30, 71

EFAS European Flood Awareness System. 7, 10, 11, 15–18, 49, 50, 52, 53, 58, 59, 61, 62, 65, 90, 105, 112, 113, 128

EM-DAT UN's Emergency Events Database. 25–28, 30, 31, 71

EO Earth Observation. 13, 17, 85–90, 95, 102, 103, 106, 107

ERCC Emergency Response and Coordination Centre. 7, 15, 16, 65

GDP Gross Domestic Product. 21, 35–37

GFMS Global Flood Monitoring System. 16, 69

GloFAS Global Flood Awareness System. 7, 10, 16, 17, 65, 67, 69, 70, 76–80, 90

GNI Gross National Income. 21, 35

JRC Joint Research Centre. 109, 117

LIDAR LIght Detection And Ranging. 88, 89

NER Named Entity Recognition. 18, 32, 33, 110, 113

NLP Natural Language Processing. 17, 18, 23, 93, 95, 121

NUTS Nomenclature of territorial units for statistics. 19, 52, 53, 58, 59, 62, 111

RM Copernicus Rapid Mapping. 8, 86, 87, 102, 106

RRM Copernicus Risk and Recovery Mapping. 86

SAR Synthetic Aperture Radars. 86, 88, 89

SMDRM Social Media for Disaster Risk Management. 10, 13, 105–108, 111, 113, 114

SMFR Social Media for Flood Risk. 10, 50, 51, 53, 59, 62, 65, 112

TRMM Tropical Rainfall Measuring Mission. 16, 69

UN United Nations. 23, 25, 71

Part I

Background

Chapter 1

INTRODUCTION

1.1. Motivation

Over millennia humans have developed villages and cities near water bodies, mainly for two reasons: (i) improvements in agricultural yields due to yearly floods in fertile floodplains leaving nutrient-rich silt deposits behind; and (ii) people's desire to live near coastlines and river valleys, often on wetlands and back-filling otherwise natural flood buffers. Due to the closeness of human settlements to rivers and coasts, floods are the natural disasters with the greatest damage potential and the ones that affect the greatest number of people [98]

Under the Paris climate agreement signed on the 4th November 2016, many countries have committed to keeping global average temperature rise well below 2°C and aim to limit the increase to 1.5°C, while increasing the ability to adapt to the adverse impacts of climate change.

Riverine floods

Considerable increases in riverine floods impacts are predicted even under the most optimistic scenario of 1.5°C warming as compared to pre-industrial levels. The Asian continent and Sub-Saharan Africa are the most affected region, and will have rising shares of the global direct and indirect impacts at all analysed warming levels. In Europe, Central and Western regions will be affected even if the temperature increases remain close to the lower estimates. For instance, flood peaks with magnitude as it happens once in 100 years, called 'return period'[1], are

[1]The return period of an event is the time span it would take to observe one such event on

projected to double in frequency within the next three decades [3]. Long term losses could even exceed direct damages, due to the increased persistent effects on the economy.

Coastal floods

Global warming is also expected to drive increasing extreme sea levels (ESLs) and storm-surge flood risk along the worlds coastlines. Projections of ESLs for the period between 2000 and 2100 show a very likely increase of 34–76 cm under a moderate-emission-mitigation-policy scenario and of 58–172 cm under a business as usual scenario. By the end of this century this applies to most coastlines around the world, implying unprecedented flood risk levels unless timely adaptation measures are taken. Areas like the North Sea on the German coast, as well as parts of East Japan, China, North Vietnam and many of the South Pacific Small Island Developing States are projected to experience the highest increase in the median ESL100 exceeding 1 m under the highest greenhouse gas emission toward the end of the century. The increase in ESLs is weaker along the coasts of the Baltic Sea, where glacial isostatic adjustment[2] results in a relative sea-level fall that counter-balances and in some cases reverses the rise in mean sea level and climate extremes [101].

Flash floods and urban floods

Climate change is expected to increase the regime of extreme precipitation. Frequency changes reveal a coherent spatial pattern with increasing trends being detected in large parts of Eurasia, North Australia, and the Midwestern United States. Globally, over the last decade of the studied period have shown a 7% increment in extreme events than the expected number, although findings report that changes in magnitude are not in general correlated with changes in frequency [70]. More rain will increase likelihood of flash floods of short duration but high intensity along with urban floods due to high density populated areas unable to cope with heavy rainfall.

---

expectation.

[2]Glacial isostatic adjustment (GIA) describes the adjustment process of the earth to an equilibrium state when loaded by ice sheets

1.2. Flood Risk Management in Europe

The Emergency Response and Coordination Centre (ERCC), operating within the European Commission's Civil Protection and Humanitarian Aid Operations department, was set up to support a fast and coordinated response to disasters both inside and outside Europe using resources from the countries participating in the EU Civil Protection Mechanism.[3] This centre monitors hazards and risks, collects and analyzes real-time information on disasters, prepares plans for the deployment of experts, teams and equipment. ERCC in general coordinates the EU's disaster response efforts when a single member state cannot cope with a crisis with its own capacities.

Situational awareness: the European Flood Awareness System (EFAS)[4] and The Global Flood Awareness System (GloFAS)[5] provide real-time information and forecasts about floods to the ERCC as well as to a series of partners including national and regional hydrological services. EFAS and GloFAS are part of the Copernicus Emergency Management System (CEMS), and holds regularly updated flood-related information such as probabilistic medium-range flood forecasts (including short-range flash floods), seasonal forecasts, and impact assessments and early warnings.

Current flood hazard mapping methodologies, such as the one implemented in EFAS, have high scale spatial resolution capacity: $1km \times 1km$ resolution globally and $100m \times 100m$ resolution in Europe [17]. This combined with state-of-the-art forecasting models provide detailed information about the risk associated with a flood in terms of likelihood, magnitude, timing, and impact [16].

Due to nature of floods phenomena and its dynamics, it is important that the flood extent can be monitored during the flood peak, therefore early warnings are important to allow sufficient lead time for requesting satellite mappings of the area at risk and collecting information about impact and risk as the event develops [18]. However the forecasting systems have limitations. Due to lack of data, computational cost and uncertainty in model meteo input, many of the global flood models use parameters that are not properly calibrated against streamflow

[3] https://ec.europa.eu/echo/what/civil-protection/eu-civil-protection-mechanism_en

[4] http://www.efas.eu/

[5] http://www.globalfloods.eu/

observations. The mentioned systems cannot forecast events for small rivers (mostly flash floods) because the data used for calibrating the model are aggregated as grids of data at a low-resolution. Again, urban floods or coastal floods are not due to river flows but because of sewage system clogged and storm surges, therefore the model simulating the flow of water in rivers is not bringing high precision flood forecasts.

Impact assessment: emergency managers use flood maps based on either hydraulic models or remote sensing data. Hydraulic models require detailed digital information of the impacted area and forecasts that may not be available readily or at the desired spatial granularity. The CEMS On-Demand Mapping, which has operated since February 2015, consists of a set of information services funded by the European Commission.

On average, the minimum time needed by Copernicus Rapid Mapping (RM) service to provide crisis information after an activation request by an authorized user[6] is 24 h [103]. Due to the technical issues discussed in Chapter 6, remote sensing analysis is of limited use in urban areas to the point that these areas are commonly not analyzed and left out of the product map.

1.3. Goals

Over the past decade, social media has emerged as a relevant data source about disasters, prompting researchers from diverse areas to converge on this domain [10, 69]. Social media analysis has demonstrated the potential to provide timely, precious information about the spatial [8, 79] and temporal [100] development of a crisis, as well as supporting the identification of key disaster-related events [65].

This research work aims at studying how well and how far information elicited from social media can be used to improve operational systems for Disaster Risk Management (DRM), reducing uncertainty and providing tools for catching not-yet monitored events, as well as establishing references data-sets for future works.

The research work has been planned for answering the following research questions:

- *RQ1: Is it possible to integrate effectively social media signals with au-*

[6]EU Member States, the Participating States in the European Civil Protection Mechanism, the Commission's Directorates-General (DGs) and EU Agencies, the European External Action Service (EEAS), as well as international Humanitarian Aid organizations

thoritative data at a pan-european level where riverine flood likelihood is estimated?

The answer to this question should be a first set of experiments confirming the possibility to listening to social media based on geographical and temporal forecast likelihood. Such forecasts are available at a pan-european scale at a regional resolution. A model for the classification of tweets needs to be trained in several available languages.

- *RQ2: Is it possible to classify reliably the relevance of social media information to floods using a 'zero-shot' transfer learning ?*

  When a classifier for few spoken languages is available, multilingual embeddings are used for adding semantic context aimed at building a classifier for a new language without training data. The experiments should aim at identifying the best set of embeddings.

- *RQ3: Is it possible to identify floods worldwide independently from forecasts using knowledge from past events independently from hydrological forecasts?*

  This question aims at extending the coverage of floods not detectable from flood monitoring systems such as flash floods in small rivers , urban floods or coastal floods due to storm surge.

- *RQ4: Is it possible to dynamically define the risk and the impact of a flood in a densely inhabited area at high resolution?*

  The main purpose of this question is to research the possibility to extract valuable information for local crisis responders. While others questions could be answered confirming the events and defining its extent and severity aggregating the classified signals from social media, *RQ4* purpose is to conduct experiments using classifiers to detect categories of infrastructures (educational, transportation, energy, etc) and aggregating signals at a high resolution.

1.4. Challenges

Although the value of social media analysis in providing timely data and methods for the analysis of natural hazards has been recognized in previous

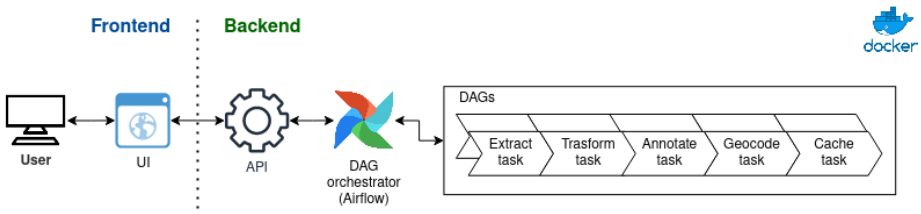


Figure 1.1: SMDRM architecture

work, comparatively much less attention has been given to how to integrate social media in a seamless, reliable way with tools for disaster risk management.

Analysis of data will be performed using machine learning models trained with data manually annotated and databases of previous events. The need of collecting data and running classification models in near-real-time lead to the development of a software framework named Social Media for Flood Risk (SMFR) which provides information from social media about flood risks and impacts associated to an event, including examples of tweets about it.

The platform could however be improved to cope with a wider geographical extent and to cope with all the type of floods. That is why, after some SMFR positive results, we created a platform with a series of containerized microservices. In the framework of the experiments for *RQ4*, we built a platform independent of the type of hazards, more focused on modularity and scalability. In the Chapter 7 we describe the creation of the platform Social Media for Disaster Risk Management (SMDRM), whose architecture is shown in Figure 1.1

Its architecture and modules are described in Chapter 7.

Since the geographical domain of EFAS and GloFAS products covers an area where population speaks more than a hundred different languages, data analysis must focus on a multilingual system.

This work aims at filling this gap by describing the integration of social media monitoring into a flood monitoring and forecasting platform, enriching hydro-meteorological information with reports from the public with a multilingual approach.

1.5. Contributions

The main contributions of this Ph.D are:

- a system that integrates social media analysis into EFAS. This integration

allows the collection of social media data to be automatically triggered by flood risk warnings determined by a hydro-meteorological model. We also describe a method for selecting relevant and representative messages and displaying them back in the interface of EFAS.

Lorini *et al.* - 2019 - «Integrating Social Media into a Pan-European Flood Awareness System:A Multilingual Approach».

The article has been accepted and presented at the 16th International Conference on Information Systems for Crisis Response and Management *ISCRAM19* in 2019 where it has been awarded as the 'Best Paper'

- a study about how the usage of non-authoritative data for disaster management provides timely information that might not be available through other means. Wikipedia, a collaboratively-produced encyclopedia, includes in-depth information about many natural disasters, and its editors are particularly good at adding information in real-time as a crisis unfolds. In this study, we focus on the most comprehensive version of Wikipedia, the English one. Wikipedia offers good coverage of disasters, particularly those having a large number of fatalities. However, by performing automatic content analysis at a global scale, we also show how the coverage of floods in Wikipedia is skewed towards rich, English-speaking countries, in particular the US and Canada. We also note how coverage of floods in countries with the lowest income is substantially lower than the coverage of floods in middle-income countries. These results have implications for analysts and systems using Wikipedia as an information source about disasters.

  Lorini *et al.* - 2020 - «Uneven Coverage of Natural Disasters in Wikipedia: The Case of Floods».

  The article has been accepted and presented at the 17th International Conference on Information Systems for Crisis Response and Management *ISCRAM20* in 2020.

- we explore the possibility of having an entirely independent flood monitoring system which is based completely on social media, and which is completely self-activated. Social media can be used for disaster risk reduction as a complement to traditional information sources, and the literature has suggested numerous ways to achieve this. In the case of floods,

for instance, data collection from social media can be triggered by a severe weather forecast and/or a flood prediction. This independence and self-activation would bring increased robustness, as the system would not depend on other mechanisms for forecasting. We observe that social media can indeed help in the early detection of some flood events that would otherwise not be detected until later, albeit at the cost of many false positives. Overall, our experiments suggest that social media signals should only be used to complement existing monitoring systems, and we provide various explanations to support this argument.

Lorini *et al.* - 2020 - «Social Media Alerts can Improve, but not Replace Hydrological Models for Forecasting Floods».

The article has been accepted and presented at the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology *WI-IAT'20*.

- a paper that summarizes key opportunities and challenges identified during the workshop 'Social Media for Disaster Risk Management: Researchers Meet Practitioners' which took place online in November 2020. It constitutes a work-in-progress towards identifying new directions for research and development of systems that can better serve the information needs of emergency managers.

  Practitioners widely recognize the potential of accessing timely information from social media. Nevertheless, the discussion outlined some critical challenges for improving its adoption during crises. In particular, validating such information and integrating it with authoritative information and into more traditional information systems for emergency managers requires further work, and the negative impacts of misinformation and disinformation need to be prevented.

  Lorini *et al.* - 2021 - «Social Media for Emergency Management: Opportunities and Challenges at the Intersection of Research and Practice».

  The article has been accepted and presented at the 18th International Conference on Information Systems for Crisis Response and Management *ISCRAM21* in 2021.

- a study aiming to determine how social media information can reduce the inherent uncertainty of the information in the immediate aftermath of an

urban flood event. Before urban flooding actually happens, weather forecasts with varying degrees of precision are available to emergency managers. In the aftermath of the event, authoritative information including Earth Observation (EO) data can be used to estimate precisely the flood extent, possibly after several hours. Specifically, the study investigates how to collect relevant social media images and to interpolate such data in order to create a map.

The premise of the study is that social media platforms, when combined with digital surface models, can provide control points for creating a reliable near real-time estimate of the flood extent. In the study, we compared a flood extent map derived from social media with that derived from authoritative altimetry data during one of the worst floods to hit Venice, which occurred in November 2019.

The results of the experiments show a good overall accuracy using several digital surface models. Given the global coverage of such models and the low resources required, we think the methodology proposed could be beneficial for emergency managers. Specifically, we describe how a flood extent map can be made available within 24 h, or even less, after urban flooding strikes a densely inhabited area, where data generated by the public are available.

Lorini *et al.* - 2022 - «Venice Was Flooding... One Tweet at a Time.»

The article has been accepted at the 25th ACM Conference On Computer-Supported Cooperative Work And Social Computing *CSCW22* in 2022.

- the platform SMDRM, which is a software platform that streamlines the processing of text and images extracted from Twitter in near real-time during a specific event. Social media has been described as a mechanism for understanding a situation using information spread across many minds, i.e., a form of distributed cognition [34]. Gaining situational awareness in a disaster is critical and time-sensitive. Social media provides a vast data source that might help improve response in the early hours and days of a crisis. Our article depicts the structure of the platform for the operationalisation of the data processing in the specific domain of disaster management.

  Lorini *et al.* - 2022 - «SMDRM: A Platform to Analyze Social Media for Disaster Risk Management in Near Real Time».

The article has been accepted and presented at the Workshop on Social Media for Emergency Response *SOMMER22* during the 16th International Conference on Web and Social Media *ICWSM* in 2022.

1.6. Thesis Outline

The thesis is organized in four parts as follows. In the first part we define the background information in which we introduce the goals of the research (Chapter 1) , we describe the state-of-the-art on social media analysis for crisis response (Chapter 2), and we try to gather flood information from Wikipedia (Chapter 3).

In the second part we explain how social media can be used as a mean for improving situational awareness during a flood event. We describe the implementation of a framework for the extraction, filtering and aggregation of flood related messages at a pan-european during floods response (Chapter 4). We then study the possibility of using such methodology to identify floods event without support of hydrological simulations.

In the third part of the work we describe our experiments towards impact assessment in terms of mapping (Chapter 6) and impacts on infrastructures/services/population (Chapter 7).

In part four, we conclude the thesis suggesting a path forward for a practical application of the methodologies implemented and described in the previous chapters from practitioners' perspective (Chapter 8). Finally we draw conclusions and suggest future line of research.

Chapter 2

RELATED WORK

At the beginning of our research work we gathered information about the relevant research work done on the two main topics we want to integrate; flood monitoring and social media analysis during crisis. We propose to study an integration at a continental and global scale, therefore we focused on large scale flood monitoring tools and multilingual analysis of social media messages.

2.1. Flood Detection and Flood Forecasting at the Pan-European Scale: the European Flood Awareness System

EFAS is part of the Copernicus Emergency Management Services, and serves the ERCC and EU member states' Civil Protection agencies with forecasts of flood risks. It covers the European Union as well as several neighbouring areas that are relevant from the perspective of flood risk and EU policies (e.g., all countries of the Danube river basin). EFAS provides information based on weather forecasts[1] and hydrological ensemble predictions obtained by an hydrological model simulations, therefore subjected to uncertainty. For being exploited at its maximum potential it should offer additional supporting information that could be used for prioritizing resources and interventions. EFAS already includes products that go in that direction. For instance pre-tasking satellite mappings is triggered in advance integrating socio-economic data (population, infrastructures,

[1]EFAS inputs are based on the Ensemble Prediction System from the European Center for Medium Range Weather Forecast (ECMWF) which consists of 51 ensemble member.

economic losses fused as impact indicators) with flood magnitude and probability (likelihood)[18].

EFAS forecasts have good accuracy level which is constantly monitored[2], but Crisis are only declared, formalised and managed by the local authorities and it's up to them to decide what to do with the data provided. The research presented here helps in the near-real-time confirmation of floods for areas where a high flood risk was forecasted. The study of social media during floods has unique challenges compared with its use for other types of disasters such as earthquakes. For the latter, there are records from seismographic monitoring networks using widely accepted standards, and therefore reliable lists of georeferenced events are available [82, 74]. For flood events, however, there is no such international standard for recording and reporting information, let alone any unique identification number of these events. Instead, emergency managers report information according to their own interpretation and local guidelines. Despite these limitations, previous work has demonstrated how social media can be used to detect floods [8], with the aim of augmenting situational awareness [84].

2.2. Flood Detection and Flood Forecasting at the Global Scale

While there are many works describing tools and methodologies for flood detection at a sub-national and national scale using high-resolution models and local sensors, few systems try to monitor floods at a global scale covering areas where authoritative data are lacking. The real-time Global Flood Monitoring System (GFMS) described in Wu *et al.* (2014), uses the Tropical Rainfall Measuring Mission (TRMM) Multisatellite Precipitation Analysis (TMPA) rainfall for detecting floods through satellite images. It has been developed and implemented using a physically based hydrologic model with temporal and spatial resolution respectively 3-hours and 0.125 degrees, which translate in, more or less, $10km \times 10km$ per cell in a gridded representation of the globe.

GloFAS, together with EFAS, is part of the Copernicus Emergency Management Systems, and serves the ERCC and EU member states' Civil Protection agencies with forecasts of flood risks. GloFAS covers , with a spatial resolution ($25km \times 25km$) lower than EFAS ($5km \times 5km$), the whole globe . They provide information based on weather forecasts[3] and hydrological ensemble pre-

[2]publicly available at https://www.efas.eu/validation-and-skill-scores

[3]Inputs are based on the Ensemble Prediction System from the European Center for Medium

dictions obtained by an hydrological model simulations, therefore subjected to uncertainty. Both the global monitoring systems though, are based on models with such resolution enabling to forecast and detect only floods in big rivers and not at all small flash floods or urban floods or coastal floods yet. The aim of this research is to extend the detection of events using social media information.

Copernicus offers information services that draw from both satellite EO and in-situ (non-space) data. The CEMS support local authorities and communities needing information to develop environmental legislation and policies or to take critical decisions in the event of an emergency, such as a natural disaster or humanitarian crisis. The Early Warning Systems (EWS) and On-Demand Mapping components of the CEMS produce flood hazard maps that have been developed using hydrological and hydrodynamic models, driven by the climatological data of the European and Global Flood Awareness Systems (EFAS [93] and GloFAS [4]). All maps are in raster format with a grid resolution of 100 m (European-scale maps) and 30 arcseconds (global-scale maps). These maps can be used to assess the exposure of population and economic assets to river floods, and to perform flood risk assessments. Our research evaluates the integration of such maps tih signals from social media for the the confirmation of floods.

2.3. Urban Flooding

Research on hyper-resolution definition of urban flooding are rare due to their complexity and the limited numbers of models fit for that purpose. Previous works try to detail flooding risk analysis, urban flooding control, and the validation of hyper-resolution numerical models [102, 79].

In Wang *et al.* (2018) social media information is filtered and aggregated using Natural Language Processing (NLP)) and images are gathered through crowdsourcing before being processed by a computer vision system. Tweets are aggregated (summed) at sub-metropolitan area.

Restrepo-Estrada *et al.* (2018) use aggregation of tweets related to rain as proxy variable indicators of the intensity of rainfall, then the variable is fed into a model for computing discharge values. The city area is divided in sub-zones.

The research proposed here wants to study if social media could be useful to detect high resolution extent and impacts of urban floods in densely populated cities where social media and socio-economic data are expected to be abundant as more users are actively connected.

Range Weather Forecast (ECMWF) which consists of 51 ensemble member.

Multi-class classification is nowadays widely know and applied especially in NLP in tasks like Named Entity Recognition (NER) or sentiment classification [50, 48]. Some work has been done trying to define multiple classes in detailing urban flooding, but their work applies classification to images rather than text, relying on fewer data than the ones available in the immediate aftermath of an event [25, 88]. The research proposed contributions go into the direction of joining the two different efforts, applying multi-class classification to text from social media in order to define the impacts of an event in near-real-time.

Urban areas offer the possibility to research if and how social media could be used also to detect inhabitants mobility, which combined with hazard static maps and maps of critical infrastructures, could be used to define a near-real-time risk assessment to crisis responders. In Park *et al.* (2018) information from several platforms are used in combination with mobile data to identify the dynamics of residents of a city, while Botta *et al.* (2019) use Instagram posts to measure the size of crowds in specific places. The proposal would like also to try to geocode social media posts with the goal of identifying crowd at risk during an event.

2.4. Combining Authoritative and Non-Authoritative Data

A recent trend in research on social media on disaster has been to study methodologies for combining non-authoritative and authoritative data in risk assessment. The non-authoritative data are reports generated by the public, typically posts in social media platforms. The authoritative data comes from various sensors including meteorological and hydrological ones as well as physical models for creating forecasts with this data. In previous work, these data have been combined in various ways.

Musaev *et al.* (2015) describe the LITMUS platform, which collects and filters messages about landslides from various social media platforms and geolocates them to merge reported events with data from physical sensors referring to the same location.

Restrepo-Estrada *et al.* (2018) use a transformation function for creating a proxy variable for rainfall by analyzing keywords-filtered geo-located social media messages and rainfall measurements from authoritative sources. The proxy variable is incorporated in a hydrological model for stream-flow estimation.

Our work differs from the previous ones in important ways. First, we do not filter social media posts by flood-related keywords, but rather according to the location of a possible flood based on forecasts from the EFAS system. Second,

we aggregate data using a geocode standard for referencing the subdivisions of countries for statistical purposes, known as Nomenclature of territorial units for statistics (NUTS) [20]. Its granularity is identified by levels, the higher the level, the higher the granularity. In our work we used level 2 (NUTS-2) as the main subdivision. Third, we cover a large area by automatically processing content in several languages. The goal of our system is to confirm and bring more detail to the outcomes of an hydrological model.

The potential of social media for situational awareness during emergencies has been studied by several researchers [10] [86]. Research has also been carried out on how emergency managers could use the information shared by witnesses to plan relief operations [76] assessing impacts at an early stage. Text and images shared on Twitter have been recognized as containing important information pertinent to humanitarian response [1]. Recent studies have shown encouraging research results related to the use of social media sensors to map flood extent. Brouwer *et al.* (2017) presented a methodology for detecting riverine flood extent using locations derived from Twitter and a normalized digital terrain model. Hydrologically connected tweets are interpolated according to a drainage-normalized representation of the topography. This approach applies to floods driven by an overflow of water from riverbeds, and the focus is on directions of flow in nearby areas rather than on urban floods. Around the same time, Rosser *et al.* (2017) presented a work to estimate flood extent based on a Bayesian model fusing remote sensing, social media and topographic data sources. The method uses geocoded photographs sourced from social media Flickr, optical remote sensing and high-resolution terrain mapping to estimate the probability of flooding through weights-of-evidence analysis. The results demonstrate that the incorporation of multiple sources of data can aid the prediction of flood extents. Their work does not consider temporal aspects of the data within the modelling process, as the case study involved a prolonged flooding event. Heavy rainfall is often the main driver for urban floods, which can happen where there are no rivers, or the flood can be a combination of events such as blockage of the sewage system or coastal floods. Our work aims to map flood extent regardless of the driver of the event.

2.5. Multilingual classification of social media postings

The framework upon which the projects run is designed to work across multiple languages. The main processing done to messages is to determine whether

they are relevant to flood risks/impacts or not. This is done through supervised classification, which requires labeled data. However, to work across multiple languages in practice requires to be able to classify messages in languages for which there may not be labeled data yet. Implementations applied to natural disasters have been explored in the past [63, 46, 45, 57, 73, 13].

Previous works, such as Li *et al.* (2018) has shown how an approach based on embeddings works better than a simpler method based on bag-of-words when generalization is critical, including the case relevant to this study, which requires generalizing across languages.

Our research aims at evaluating and demonstrating how the classification task can be transferred across multiple languages for which no semantic resources are available, leveraging on embeddings for multilingual modelling.

Recently, considerable steps have been made towards the possibility of knowledge transferring without parallel data, also known as 'zero-shot' transfer learning, in fields such as machine translation [41, 28] but also in multilingual classification [6, 15].

Our research compares several different ways in which embeddings can be used to perform multilingual classification: using language-agnostic word embeddings learnt from a multilingual corpus [73], using multilingual word embeddings that are aligned across languages [13], using pre-trained encoders trained with sentence-embeddings as described in Artetxe and Schwenk (2018) or fine-tuning pre-trained encoders [15]. Previous research also proposed to go beyond the mere use of text. Imran *et al.* (2020 use both text and image modalities of social media data and fuse them to learn a joint representation using deep learning techniques. Specifically, they utilize convolutional neural networks to define a multimodal deep learning architecture with a modality-agnostic shared representation with good results.

Chapter 3

FLOODS AND USER-GENERATED DATA

We wanted to study Wikipedia as we can see it as a permanent social network. Information extracted can be used to carry out experiments with past recorded events from authoritative data sources. In this chapter, we estimate the coverage of floods in Wikipedia along many variables, including Gross Domestic Product (GDP), Gross National Income (GNI), geographic location, number of English speakers, fatalities, and various indices describing the level of vulnerability of a country. Addressing flaws and exposing biases can help the research community think about possible countermeasures that can lead to a set of best practices for Wikipedia or publishing research leveraging Wikipedia data, or others social networks.

3.1. Introduction

During the past decade, water-related disasters, such as floods, droughts, storm surges, cyclones, convective storms and tsunamis, accounted for 90% of all disasters in terms of the number of people affected and among them, 50% were flood events [31].

Unaddressed vulnerabilities, rising population, intertwined natural events, continue to be the main critical factors for loss of life, disrupting livelihoods and fueling new displacement. A previous analysis estimated that people in the least developed countries are, on average, six times more likely to be injured, lose their home, be displaced or evacuated, or require emergency assistance, than those in

high-income countries [99].

Death tolls and economic losses from natural hazards are expected to rise in many parts of the world. Countries with higher income levels show lower human vulnerability and the high number of people exposed translates into lower mortality compared to developing countries [19]. An analysis of vulnerability at a global scale, integrating population and economic dynamics with one of the most comprehensive natural disaster loss databases, show that there is still a considerable climate hazard vulnerability gap between poorer and wealthier countries [23].

Wikipedia, founded in 2001, has come a long way, becoming over the years one of the primary sources of encyclopedic information worldwide. In fact, during 2018 alone, the English Wikipedia had over 108 billions article views. It accounts for approximately 45% of all page views on Wikimedia projects in this period (237 billions).[1]

One of Wikimedia's goal is sharing knowledge, and an extensive international base of editors is a crucial element in providing information in several languages. Even if a useful, ethical code for Wikipedians can guide editing towards styles of practice that best support the Wikipedia mission[2], when editors mix personal interests with the goals of the Wikipedia community as a whole, they make choices that can affect the articles they create and edit [33]. Although collaborative editing fulfils the objective of sharing information, it can introduce biases that are apparent when Wikipedia is used as a reference data set for a specific topic, such as natural science research.

In recent years, researchers have placed much effort into studying how to extract meaningful information for crisis management from social media and collaborative sources [51, 37, 68], but biases in these sources are rarely evaluated.

In a seminal paper, Galtung and Ruge (1965) showed that pieces of news from 'elite' nations were more likely to be covered in foreign news reports. We find evidence of the same for the coverage of floods in the English Wikipedia, noting that floods in the wealthiest countries, particularly floods in the US, are more likely to appear in Wikipedia than floods in the poorest countries.

We think that Wikipedia is a valuable source of free data, and it could be ben-

[1] `https://stats.wikimedia.org/v2`

[2] `https://en.wikipedia.org/wiki/Wikipedia:Ethical_Code_for_Wikipedians`

eficial to researchers in the Disaster Risk Reduction field if biases are identified, measured, and mitigated. Our main contributions are:

- We establish a validated reference set of events tracked by several independent organizations, with support from hydrologists. Some organizations collect data about floods for different purposes, from insurance to sustainable development goals set by the United Nations (UN). Their effort is to collect floods data on a global scale. We compare and collate the different data sources.

- We match verified events with Wikipedia entries. We analyze three methodologies for matching verified events with Wikipedia's text in terms of location and temporal references. In our work, a particular effort has been made to geo-locate Wikipedia entry candidates since we wanted to identify news reporting information about an event and to exclude generic collections of unspecified events.

The remainder of this chapter is organized as follows: the next section presents related work; then, the third section describes the methods for establishing verified ground truth information, for matching Wikipedia data with verified events and how to geo-locate them. Finally, we present experimental results, including a case study, followed by our conclusions and future work.

3.2. Related Work

Wikipedia has been used as a data source to study *sustainable development* and for *Disaster Risk Reduction.* For instance, it has been recently used as a source of data to estimate indicators at very high spatial resolution leveraging recent advances in NLP by extracting information from free-text articles [92]. In their work, the spatial distribution of the articles and meta-data extracted from its text, combined with other data such as night-light satellite images, are used to improve the prediction of socio-economic indicators of poverty, as measured by ground truth survey data collected by the World Bank.

In previous work, researchers used Wikipedia for detecting and monitoring natural disasters [94] leveraging interlinks between versions of the same article in different languages and inbound/outbound redirects to other similar articles. The methodology proposed in their paper consists in creating and maintaining a list of articles related to natural disasters, scanning Wikipedia entries and subsequently

checking if edits happen on an article in the list, assuming a new event reported would impact the monitoring-list.

Considering that Wikipedia is being used as a source for data analysis, our work aims at identifying potential biases in Wikipedia coverage of natural events, specifically floods.

Wikipedia exhibits a substantial amount of self-focus, in the sense that editors in each language-specific Wikipedia tend to write about topics that are of interest to their community and not others [30]. A country-based analysis of Wikipedia shows that geotagged articles (i.e., articles referring to specific locations) concentrate in only a few countries, and this concentration can be explained in no small extent with variables such as population, number of broadband Internet connections, and number of edits emanating from each country [26].

A comparison of Wikipedia with the Global Terrorism Database[3] in 2015 shows that Wikipedia covered about 78% of attacks and almost all of the terrorism-related deaths in Western countries, but only 28% of those in other countries [85].

Also, Wikipedia suffers from a cultural gap that favours entries written in English and especially, those referring to the United States of America (USA) which are the longest and best-referenced ones [9]. Tobler's law for geography claims that similarity decreases with distance [96]. According to this law, those events happening close to English speaking countries should be considered more familiar to Wikipedia editors and therefore, better covered than those happening in distant places. There are also urban/rural biases, with Wikipedia coverage of rural areas being systematically inferior in quality [40].

Becoming a source on current news events was not part of the original mission of Wikipedia, but currently, the most visited and edited articles are about current events [44]. Wikipedia has transitioned into a source that incorporates significant news work [42].

When it comes to history, Wikipedia narratives are biased towards recent events and those happening in Western Europe [89]. Partially because of this, there is an explicit Wikipedia policy against 'recentism'.[4]

Regarding coverage of natural disasters, a study on the Tōhoku catastrophes showed that activity on Wikipedia concentrated on the day of the earthquake, but there was intense editing activity for several days [43]. A similar pattern of

[3] https://www.start.umd.edu/data-tools/global-terrorism-database-gtd

[4] https://en.wikipedia.org/wiki/Wikipedia:Recentism

intense activity close to the events was observed in the 2011 Arab spring [22]. Most of these event-centric articles are written as the event unfolds [64] and indeed, spikes in editing activity can be used for detecting new crisis events [94].

Our work is focused on natural disasters at a global scale for events happening over more than three years. Therefore, our experiments widen the previous analysis of biases, including a set of socio-economic risk indicators concerning natural hazards.

3.3. Methods

This work aims to analyze the coverage of floods in Wikipedia. Development and relief agencies have long recognized the crucial role played by data and information from previous events in mitigating the impacts of disasters on vulnerable populations. Due to the complexity of collecting reliable information, there is still no international consensus regarding best practices for defining critical aspects of an event such as starting date, duration or number of fatalities.

To carry out our experiments, we selected data source which included validated information from international relief agencies or local governments worldwide to cover all the events that could have been detected on the social networks so Precision and Recall could be computed against a complete validated dataset. That is why we consider three of the most comprehensive databases documenting floods that are commonly used by the hydrology science for reference[104]:

- **Floodlist**[5], funded by the EU Space program Copernicus [6] program, it reports on all the major flood events from around the world. Floodlist includes articles on flood-related issues such as warning systems, mitigation and control, flood recovery, flood damage repair and restoration, as well as flood insurance. The reports and articles also include information about the extraordinary humanitarian, aid and relief efforts made in the aftermath of many flood disasters.

- The **UN's Emergency Events Database (EM-DAT)**[7] contains information from various sources, including UN agencies, non-governmental organizations, insurance companies, research institutes and press agencies. Data from UN agencies, governments, and the International Federation of

[5] `https://floodlist.com/`
[6] `https://www.copernicus.eu/en`
[7] `https://www.emdat.be`

Red Cross and Red Crescent Societies have priority. This choice is not only a reflection of the quality or value of the data, but it also reflects the fact that most reporting sources do not cover all disasters or have political limitations that could affect the figures. The entries are reviewed continuously for inconsistencies, redundancy, and incompleteness.

- The **Dartmouth Flood Observatory (DFO)**[8], based at the University of Colorado, maintains the Global Active Archive of Large Flood Events derived from news, governmental, instrumental, and remote sensing sources. The archive is 'active' because current events are added immediately. Each entry in the archive and related 'area affected' map outline represents a discrete flood event. The listing is comprehensive and global in scope.

We also looked at other reliable sources such as the Copernicus Emergency Manamagent Services based on requests for satellite images acquisition for emergency response and risk and recovery maps. We found that official requests were issued only when the national authorities could not cope with the disaster on their own, resulting in only one-tenth of events recorded by the other sources, most of which were redundant.

As shown in Table 3.1, none of the three selected databases is complete, and some events recorded in one database are not in the others. Hence, we merge multiple databases into a single dataset. Our data begins on 2016-02-29 because this is the earliest date for which the three datasets contain information.

Year	Total floods	Floodlist	EM-DAT	DFO
2016	261	191	169	99
2017	322	220	215	117
2018	394	306	191	157
2019	125	96	74	42
Total	1102	813	415	649

Table 3.1: Number of events recorded per year in chosen data sources between 2016-02-29 and 2019-05-20

[8] `https://http://floodobservatory.colorado.edu/`

Criteria and definitions of events
It is essential to assess criteria for event recording and limitations of the several data sources before homogeneously merging their data.

- Floodlist includes articles on flood-related issues such as warning systems, mitigation and control, flood recovery, flood damage repair and restoration, as well as flood insurance. We decided to leave out news items with 'landslides' as the only tag, while we ingested all the other news items as we think they were mostly relevant to floods.

- In EM-DAT, for a disaster to be entered into the database at least one of the following criteria must be fulfilled: (i) ten (10) or more people reported killed, (ii) one hundred (100) or more people reported affected, (iii) a declaration of a state of emergency, and/or (iv) a call for international assistance. EM-DAT provides geographical, temporal, human and economic information on disasters at the country-aggregated level. When the same disaster affects several countries, EM-DAT enters several country-level disasters into the database with the same identifier. From all the EM-DAT database, we consider only events labelled with 'flood' or 'storm' as primary disaster type.

- DFO derives from a wide variety of news and governmental sources. The quality and quantity of information available about a particular flood are not always in proportion to its actual magnitude, and the intensity of news coverage varies from nation to nation. DFO creates a record for any flood that appears to be 'large, with, for example, significant damage to structures or agriculture, long (decades) reported intervals since the last similar event, and/or fatalities. Deaths and damage estimates for tropical storms are totals from all causes, but tropical storms without significant river flooding are not included. No filter is applied to information as we assumed all the news items were relevant to floods.

Since our ground truth information's main purpose is to support the analysis of the coverage of events in Wikipedia, we opted for a rather inclusive definition of flood and included events associated to heavy rainfall, which is the first driver of an overflow of water in river channels but also in coastal and urban areas. The merged database contains information from different sources, trying to avoid duplicates. We aggregated events at the national level, and when an event affected

more than one country, we insert a record for each country with the same dates. For those events for which a data source did not indicate the end date, we assumed it was three days after the starting date of the event. We choose three because it is the median value of the duration of the floods in our dataset. An example record contains the following information:

- start_date: year, month, day
- end_date: year, month, day
- country: name of the country
- affected: string from the source about population affected
- fatalities: number of deaths associated with the event
- location_source1: location from location_source1
- location_source2: location from location_source2
- location_source3: location from location_source3
- identifier: list of id from sources
- disaster_type: i.e Storm, Flash Flood, Flood
- in_emdat: True/False
- in_dartmouth: True/False
- in_floodlist: True/False

The resulting ground truth dataset contained 2295 floods. However, there were still many duplicate items needing to be consolidated.

The criteria for describing an event vary according to the source. Starting and ending dates are difficult to establish and there is no agreed methodology for reporting about duration and impacts of floods among the Disaster Risk Reduction community.

For instance, a flood in Angola on March 2016 was reported by Floodlist as spanning the period 2016-03-05 to 2016-03-07, by EM-DAT as happening from 2016-03-01 to 2016-03-10, and by DFO as occurring from 2016-03-01 to 2016-03-10.

To avoid duplicates, when two or more events from the data sources overlapped in time and country, the earliest starting date was selected as starting date for the event and the latest ending date as the ending date. This choice also means that two events happening at an overlapping time in two different locations of the same country will be considered as duplicates and merged. The aggregation was because the DFO dataset can locate an event only at the national level. We normalized country names of each source to facilitate the merging process.

After the consolidation process, the dataset consisted of 1102 flood events. Figure 3.1 shows a Venn diagram illustrating the intersection between and among our three data sources. The intersecting areas are consistent and represent the majority of events, meaning that more than one source identified such floods. We asked experts in the field of DRR to analyze a sample of twenty records that

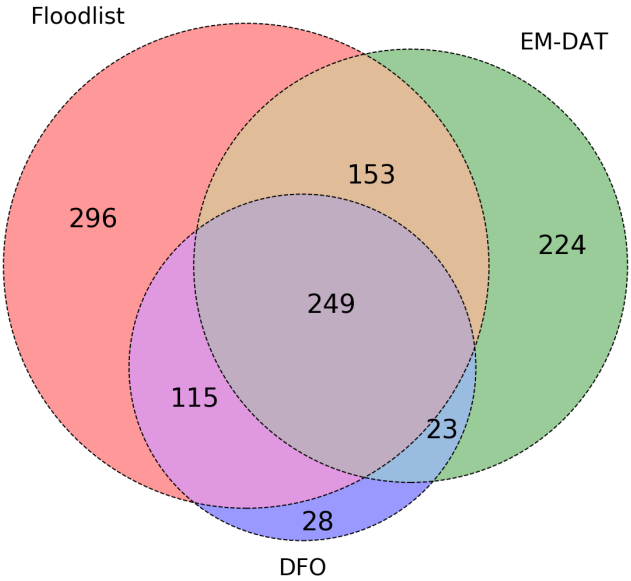
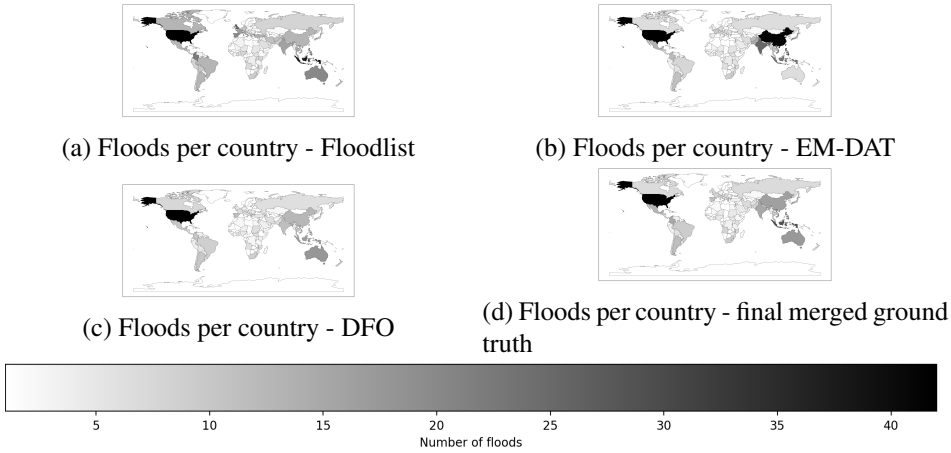


Figure 3.1: Venn diagram representing the intersection of floods coming from the data sources

appeared only in one data source. They convened that:

- Unique records from Floodlist were mostly due to the inclusion of land-



(a) Floods per country - Floodlist

(b) Floods per country - EM-DAT

(c) Floods per country - DFO

(d) Floods per country - final merged ground truth

Figure 3.2: Floods per country (heat maps) of the three data sources for ground truth information and the final merged dataset

slide associated with storms and episodes of heavy rain which were excluded from the other two data sources because not defined by their criteria as a flood.

- Unique records from EM-DAT were due to the inclusion of convective storms that lead to wind storms or sand storms which were excluded from the other two data sources because not associated with a flood.
- Unique records from DFO were mainly due to a country attribution different than the other sources in case of transborder events.

In light of the analysis of the data sources, we decided to conduct the experiments using the 458 events located in the intersecting areas, assuming that we can safely consider floods recorded by two or three data sources.

Figure 3.2 shows the geographical distribution of events recorded in each data source and the final merged result

To further evaluate potential biases in the distribution of events across data sources we compared the ground truth information with data from Munich RE's NatCatSERVICE[9], one of the most comprehensive natural disaster databases available, which primary interest is to assess insured losses. While the number

[9] `https://natcatservice.munichre.com/`

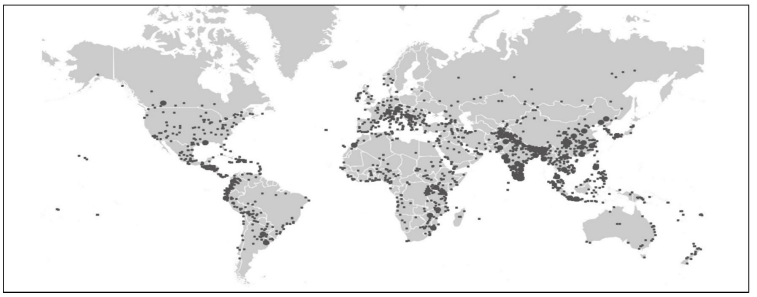


Figure 3.3: hydrological events listed in NatCatSERVICE for the period 2013-2018

of events registered seems to confirm what described by our ground truth dataset (385 hydrological events in 2018, 340 in 2017), the geographical distribution of events over the globe shown in Figure 3.3 indicates a different distribution of events. The discrepancies emerged can be explained in part by the fact that the majority of events comes from Floodlist, which also records storms associated with heavy rainfall. Such events fall into a different category of events (meteorological) in the NatCatSERVICE data.

One might think that our ground truth information could be biased by the coverage of events in wealthier countries where access to digital information is extensive and where English is the predominant language spoken. Nevertheless, the EM-DAT distribution shown in Figure 3.2b is similar to the one of NatCatSERVICE.

Finding floods in wikipedia

Initially, a keyword-based search was done on articles from the English Wikipedia covering the period for which we collected ground truth information. To do this we used a public snapshot[10]. We defined a set of keywords ('flood', 'floods', 'flooding', 'flooded', 'inundation') as representative for a potential flood event identification in a sentence. Therefore, we scanned for text containing any of the keywords within the sentences extracted from Wikipedia. If we found any of them, the sentence was stored with its corresponding title and

[10] `https://dumps.wikimedia.org` of the full English Wikipedia (containing around 6M articles), generated on May 20th, 2019.

paragraph as a 'candidate' sentence. In the case of articles whose title contained any of the keywords, such as 'Floods in the United States: 2001–present', all sentences were considered as candidates since we assumed that the content was about floods.

Of all the sentences filtered with the mentioned keywords, we selected only the ones directly linked to flooding. To ensure that, we applied to candidates a classifier (Lorini et al. 2019) expressing the probability of a text to be relevant to a flood. We selected only the ones with a probability higher than 40%. The Precision of this step was 83%, computed over a sample of two hundred candidate sentences extracted from Wikipedia.

Selected articles needed to satisfy further criteria before being checked against ground truth information. Only the sentences containing information about Country and Time-span of the event(s) were considered potential candidates. For extracting a date or location mentions, we used a Named Entity Recognition (NER) library named spaCy[11] on every title and sentence of the candidates. Subsequently, all the potential candidates were parsed to extract timestamps and countries. We created multiple candidates sentences in case spaCy returned multiple placenames.

For parsing string identified as dates, we used the datefinder[12] library which can convert strings into structured objects. It can also extract part of a date, such as 'early June'. Since sometimes a specific year was not explicitly mentioned in a sentence but could be guessed from the context, we defined the following heuristics for extrapolating the year found elsewhere in a text:

- If there was only one year within a sentence, we could assume that the same year applies to every incomplete date entity in that sentence.
- If there was no year in the sentence and only one year in its whole paragraph, we could assume that the same year applies to that sentence.
- If there was no year in the sentence and only one year in the title of the article containing the sentence, it could be assumed that the same year applies to that sentence.

The heuristics used for associating place names to their respective country was a cascade of the following heuristics:

[11] `https://spacy.io`

[12] `https://datefinder.readthedocs.io`

1. Wikidata: we searched for the placename identified by spaCy in Wikidata. If the entity returned has a corresponding page in the English Wikipedia, the country returned by the query is associated with the candidate sentence.

2. Nominatim: we searched on Nominatim[13] the place names that were not associated to a country after the first step. The query used the public Nominatim API and the country associated with the place name was the most 'important'[14] result returned.

3. Mordecai:[15] the sentences and titles not associated with any country in the previous steps were then processed using Mordecai for inferring a country from the text.

Tables 3.2 shows the results of the application of our methodology on a set of sentences. In some cases, the NER library could not find any placename; sometimes, the placename did not lead to the identification of the related country and in other cases, we could extract a country name. Finally, we discarded candidate sentences for which we could not find a country and a time reference.

Matching wikipedia candidates and ground truth information
The last part of the matching process was determining if the selected Wikipedia sentences were identifying an entry in the ground truth database. We defined three methods for identifying matching records. Here they are listed from the most strict to the laxest:

- Country and Year-Month-Days matching
  A Wikipedia candidate matches an event in the ground truth database if they link to the same country name and the date in the title or sentence of the candidate is within the time range[start_date, end_date+5 days] of the ground truth entry.

  For instance, the sentence 'On April 13, reportedly 12 people counted were killed by Rainstorm and Flash flooding in KPK and Balochistan.' matches the flood happening in Pakistan between 2019-04-13 and 2019-04-18.

[13] `https://nominatim.openstreetmap.org/`

[14] The results have a value that represents the importance of the location according to the number of citations in Wikipedia.

[15] `https://github.com/openeventdata/mordecai`

Sentence	Location entities	Countries
'The 2009 West Africa floods are a natural disaster that began in June 2009 as a consequence of exceptionally heavy seasonal rainfall in large areas of West Africa'	West Africa	None
'In the Tiquicheo Municipality, 10 houses flooded after a river near the city overflowed its banks'	the Tiquicheo Municipality	None
'The town of Poldokhtar in Lorestan Province was engulfed by flood water.'	Poldokhtar, Lorestan Province	None
'2015 Southeast Africa floods'	None	None
'New Orleans Outfall Canals'	None	None
'Serious flooding was also reported in Greenwich, Woolwich and other locations further downriver, causing major property damage.'	Greenwich, Woolwich	United Kingdom
'In July 2012, heavy torrential rains caused floods in Kyushu, Japan, leaving 32 people dead or missing.'	Kyushu, Japan	Japan
'In Antu County, 70 homes in one village were destroyed by flooding, a mountain valley was submerged by floods 20 m deep, forcing 570 families to evacuate.'	Antu County	China

Table 3.2: Examples of successful and unsuccessful attempts at location inference from text

- Country and Month-Year matching
  A Wikipedia candidate matches an event in the ground truth database if they link to the same country name and the month in the candidate sentence or title is overlapping with the time range[start_date, end_date] of the ground truth entry.

  For instance, the sentence 'In August 2018, the region yet again experienced record-breaking flooding in valley towns such as Coon Valley, Wisconsin, La Farge, Wisconsin and Viola, Wisconsin.' matches the flood happening between 2018-08-20 and 2019-08-22 in the USA.

After performing the matching for each of these pairing methods, we evaluated the hits manually for events covering three different months.

We define the **Precision** of our methodology as the fraction of matched candidates that are describing an event enlisted in the ground truth dataset. We can think about Precision as the answer to the question *How many Wikipedia matched candidates are a flood recorded in the ground truth dataset?*

We define the **Recall** of our methodology as the fraction of ground truth events that are identified by the matched candidates. We can think about Recall as the answer to the question *How many floods in the ground truth dataset are matched by Wikipedia candidates?*

Results of Precision and Recall evaluated manually over a sample of three

Period	Floods in ground truth	Precision(%)	Recall (%)
November 2016	18	66.67	16.67
September 2017	20	66.67	15.00
June 2018	26	88.89	34.62

Table 3.3: Metrics for country and year-month-days matching

Period	Floods in ground truth	Precision(%)	Recall (%)
November 2016	18	66.67	16.67
September 2017	20	50.00	20.00
June 2018	26	53.33	57.68

Table 3.4: Metrics for country and year-month matching

months from our consolidated dataset are shown in Tables 3.3 and 3.4. For the identification of a correlation between socio-economic indicators and flood coverage in Wikipedia, we opted for the matching method using Country and Year-Month-Days because a higher Precision implies that more matches are relevant, thus better support our analysis.

3.4. Results

We will use the term **Hit Rate** to refer to the percentage of matches between the Wikipedia articles and events in the ground truth. A Hit Rate of 100 means that all the floods representing a set of events in the ground truth database matched some Wikipedia candidates. A Hit Rate equal to 0 represents no coverage in the English Wikipedia for any flood of the set of events analyzed.

Our research analyzed how articles in English Wikipedia covered the floods reported worldwide in our ground truth database. We analyzed several socio-economic variables to see whether they correlate with floods coverage. These variables are GDP per capita, GNI per capita, country, continent, date, fatalities, number of English speakers and vulnerability index.

Figure 3.4 shows the top twenty countries ordered by the number of floods in the ground truth dataset and their respective hit rate. Among the countries with the highest number of floods, the United States shows a hit-rate two times higher

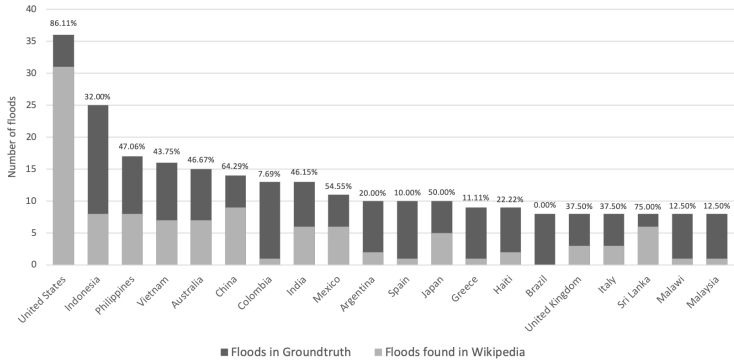


Figure 3.4: Top 20 countries ordered by number of floods in ground truth

than the ones of all the others.

Figure 3.5 shows the top twenty countries sorted by the hit-rate and their coverage in the ground truth database. Floods events in the USA and Canada are reported most frequently on Wikipedia English than anywhere else in the world. The language can be only a partial explanation because for floods in Australia the hit-rate is half and lower than other non-English-speaking countries.

Table 3.5 and Figure 3.6 show the ratio between floods recorded in the ground truth database and the floods detected in Wikipedia aggregated by continent. Since floods are geophysical event, this aggregation offer a comparison between similar Areas' extension.

Although most events happened in Asia, floods in North America have been reported more frequently.

In order to deepen our analysis, we divided the countries into six groups according to their Gross Domestic Product per capita in US Dollars, following the classification set by the World Bank for this indicator[16]:

- Low income: GDP per capita $<$ \$812
- Low middle income: \$812 $\leq$ GDP per capita $<$ \$2,218
- Middle income: \$2,218 $\leq$ GDP per capita $<$ \$5,484
- Upper middle income: \$5,484 $\leq$ GDP per capita $<$ \$9,200

[16]urlhttps://data.worldbank.org/indicator/ny.gdp.pcap.cd

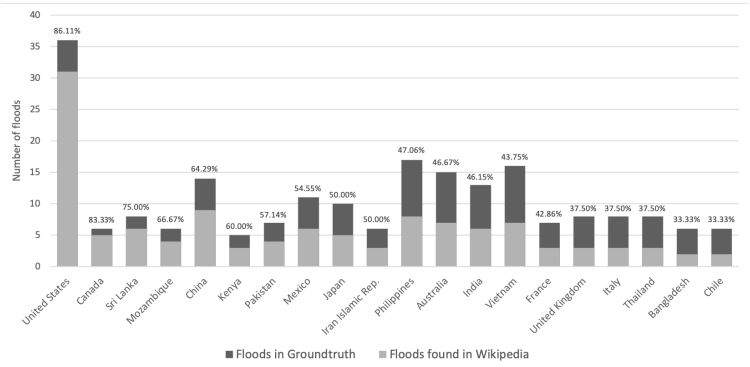


Figure 3.5: Top 20 countries with at least five floods in ground truth dataset ordered by hit rate

	Floods in ground truth	Floods in Wikipedia	Hit rate (%)
Asia	194	73	37.63
North America	106	52	49.06
Africa	96	21	21.88
Europe	85	18	21.18
South America	57	6	10.53
Oceania	27	8	29.63

Table 3.5: Percentage of ground truth floods matched by Wikipedia: per Continent

- High income: \$9,200 $\leq$ GDP per capita $<$ \$44,714
- Very high income: GDP per capita $\geq$ \$44,714

Table 3.6 and Figure 3.7 show the results for each of these groups. Within the ground truth database, floods recorded for countries in the high-income group are more than in any other group. The hit-rate of most countries from lower-middle-income to high-income varies between 30% and 40%. Hit-rate is considerably different between the lowest and the highest bracket, close to 19% for the former and 65% for the latter.

We also grouped the countries into four different groups following the clas-

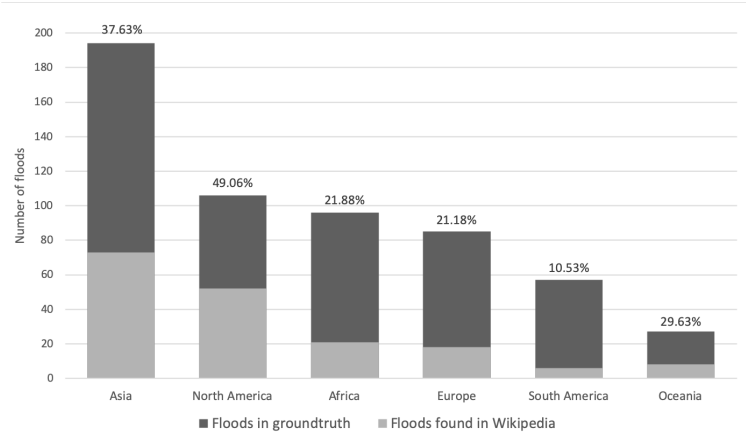


Figure 3.6: Floods for each continent and their corresponding hit rate ordered by number of floods in the ground truth

	Countries	Floods in ground truth	Floods in Wikipedia	Hit rate (%)
Low income	21	46	9	19.57
Lower middle income	37	87	29	33.33
Middle income	39	113	32	28.32
Upper middle income	27	69	14	20.29
High income	67	165	45	27.27
Very high income	28	69	47	68.12

Table 3.6: Number of floods matched per Country according to GDP per capita

sification set by the World Bank for this indicator.[17] The difference in coverage (hit rate) between high-income and low-income countries is even more evident, as shown in Figure 3.8.

INFORM Global Risk Indicators (GRI) is an open-source risk assessment tool for humanitarian crises. It can support decisions about prevention, preparedness and response [58]. We combined two socio-economic indicators that complement our previous analysis from their 2018 report:

- The *Vulnerability* indicator addresses the intrinsic predispositions of an ex-

[17] `https://data.worldbank.org/indicator/NY.GNP.PCAP.CD?view=chart`

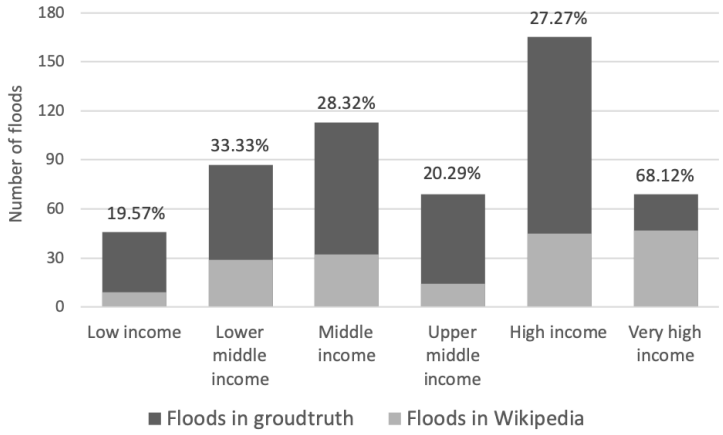


Figure 3.7: Floods for each level of GDP per capita and its corresponding hit rate

posed population to be affected or to be susceptible to the damaging effects of a hazard. So, the Vulnerability dimension represents the economic, political and social characteristics of the community that can be destabilized in case of a hazardous event. Physical vulnerability is a different matter; it is embedded into the hazard and exposure indicators.

- The *Lack of coping capacity* indicator measures the ability of a country to cope with disasters in terms of formal, organized activities and the effort of the country's government as well as the existing infrastructure, which contribute to the reduction of disaster risk.

They are both expressed on a scale of zero to ten. We combine the indicators as the square root of the product between them. We grouped the events in the ground truth database into four categories sorted by ascending value of the combined indicator. The higher is the indicator, the more vulnerable is the country. Events where the capacity to cope with a disaster is the lowest, therefore where the impact could be the highest, are less likely to be described in Wikipedia.

Our analysis also considered the percentage of English speakers[18] in a country. In Table 3.8 and Figure 3.9, we see an increase in the coverage for countries

[18] `https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population`

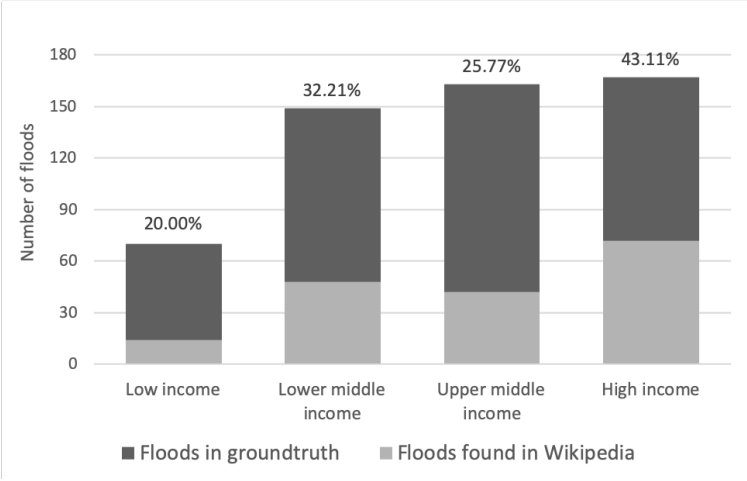


Figure 3.8: Floods for each level of GNI per capita and its corresponding hit rate

Vulnerability	Floods in ground truth	Floods in Wikipedia	Hit rate (%)
0-2 (least vulnerable)	66	21	31.82
2-4	157	60	38.22
4-6	130	43	33.08
6-8	101	31	30.69
8-10 (most vulnerable)	90	20	22.22

Table 3.7: Percentage of ground truth floods matched by Wikipedia: per INFORM indicators

with 60% or more English speakers, and then another increase for countries with 80% or more English speakers. The percentage of the English-speaking population is an indicator of the probability that an event would be described in English Wikipedia. Nevertheless, the population of a country could be related to the event coverage by English-speaking editors. We sorted the countries by ascending population and divide them into four groups containing the same number of countries each.

- Group 1: population < 754,394
- Group 2: 754,394 $\leq$ population < 6,465,513

English speakers (%)	Floods in ground truth	Floods in Wikipedia	Hit rate (%)
<20	131	36	27.48
20-40	67	19	28.36
40-60	37	14	37.84
60-80	24	9	37.50
80+	92	53	57.61

Table 3.8: Percentage of ground truth floods matched by Wikipedia: per percentage of English speakers

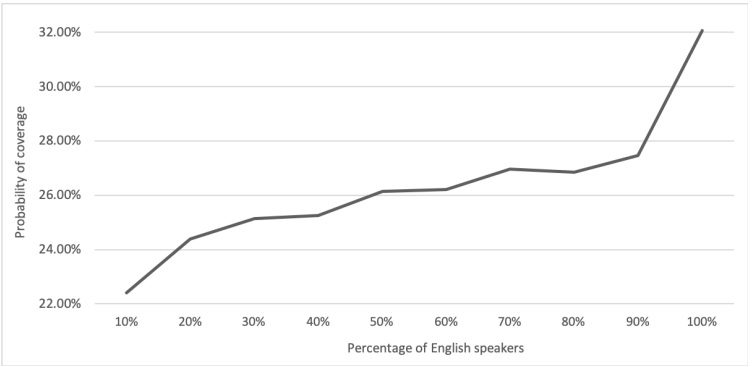


Figure 3.9: Probability of hit given the percentage of English speakers

- Group 3: 6,465,513 $\leq$ population < 24,992,369
- Group 4: 24,992,369 $\leq$ population

Indeed, as shown in Table 3.9, the hit rate is more significant for the most populated countries. In order to determine if the time of events affects the coverage of floods in Wikipedia, we analyzed the temporal distribution of the events. The method applied for evaluating hit rates showed that the relation between ground truth events and matches follow similar proportions across time, as shown in Figure 3.10. Table3.10 and Figure 3.11 shows a significant increase in the number of events that are matched by Wikipedia articles for floods in the ground truth database leading to hundred of fatalities or more. Combined with the other indicators, this could mean that only events with high impacts echoed to countries with higher English-speaking population rate and high-income to mid-income.

Country population	Floods in ground truth	Floods in Wikipedia	Hit rate (%)
G1 (smallest)	14	2	14.29
G2	60	8	13.33
G3	133	29	21.80
G4 (largest)	342	137	40.06

Table 3.9: Percentage of ground truth floods matched by Wikipedia: per population

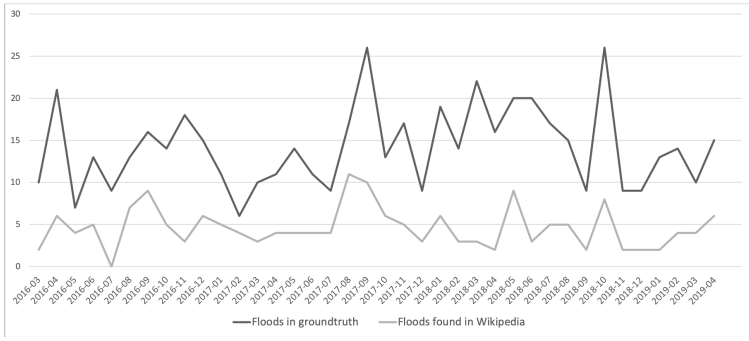


Figure 3.10: Floods for each month between 2016-03 and 2019-04

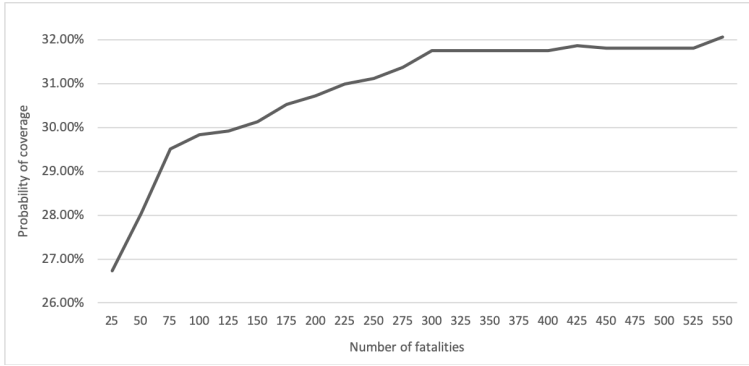


Figure 3.11: Probability of hit given the number of fatalities

3.5. A Tale of two Floods

North and central america, September 2017.
Hurricane Irma made landfall on northeast Caribbean islands during the early

Num. of fatalities	Floods in ground truth	Floods in Wikipedia	Hit rate (%)
0	164	31	18.90
1-9	184	57	30.98
10-99	173	68	39.31
100-1999	27	20	74.07

Table 3.10: Percentage of ground truth floods matched by Wikipedia:per number of fatalities

hours of 6 September, affecting Antigua and Barbuda, Anguilla, British Virgin Islands, St Barthélemy, St. Martin, the Virgin Islands and other islands in the eastern Caribbean Sea. After causing devastating damage across the Caribbean, Hurricane Irma made landfall in the Florida Keys on 10 September and worked its way north, bringing with it strong winds, storm surge and flooding rain.

Although national news agencies covered only partially the event, we can say that Irma caused between fifty and one-hundred fatalities, affecting millions of people. In our datasets, the Hurricane Irma is linked to several countries, and it produced the highest number of matches (43 total, 40 only USA), meaning many sentences on Wikipedia reported about it. Find some examples of sentences reporting Hurricane Irma in Table 3.11.

Sentence	Date	Countries
'In September 2017, Hurricane Irma storm surge caused major flooding in the downtown area of Jacksonville.'	2017-09-12	United States of America
'Moyer, Crystal (September 8, 2017). Hyatt Regency in downtown Jacksonville being evacuated.'	2017-09-08	United States of America
'People stand in a flooded street that usually serves as a farmers market, in Ouanaminthe, northeast Haiti, September 8, 2017.'	2017-09-08	Haiti
'Hurricane Irma: 10 dead in Cuba as record flooding hits northern Florida [...] September 11, 2017.'	2017-09-11	Cuba

Table 3.11: Examples of sentences reporting Hurricane Irma

Sudan, August 2018.
By August 2018 heavy rains in Sudan that had started in mid-July had caused severe flooding. As of 16 of August the floods and rain had left at least 23 people dead, over 60 injured and affected more than 70,000 people. Although the event appears in all three data sources, we could not find any match in Wikipedia applying our methodology. Either the event was not the subject of any Wikipedia article, or it was not described as accurately as other events.

Even if both events had a high number of fatalities and affected people, while the former event was widely identifiable on Wikipedia, the latter case was less (or poorly) described.

3.6. Conclusions

According to the United Nations Office for Disaster Risk Reduction, the impact of natural hazards is highest on the most marginalized populations, exacerbating inequality and further entrenching poverty. Beyond focused attribution to single events, impacts are often found to be a function of a series of associated shocks such as famine, disease and displacement that prompt disruption in multiple dimensions (e.g. livelihoods, education or labour-market) [99].

For instance, it is estimated that 35.6% of the population affected by floods in Pakistan in 2010 consequently slipped under the poverty line as a result.

The results of our analysis are consistent along several dimensions, and paint a picture in which Wikipedia's coverage is biased towards some countries, particularly those that are more industrialized and have large English-speaking populations, and against some countries, particularly low-income countries which also happen to be among the most vulnerable. This means that tools using data from social media or collaborative platforms should be carefully evaluated for biases.

Limitations
We considered only one type of event that is very prevalent globally: floods, but other types of events should be considered. We had chosen to focus on one type of event because this work is a first attempt to bring Wikipedia and its crowdsourced information into the scope of Disaster Management. Therefore the experiments and the results must be solid, reproducible and clear.

We relied on methodologies demonstrated and developed in previous work such as automated classification of text using ML models and we used consistent

exhaustive data sources.

Automated content analysis cannot replace expert annotation, but considering that the English Wikipedia contains over five million articles, it is impractical to perform this analysis manually. Some biases introduced by automated content analysis may include the usage of libraries for parsing geographical entities, which may have been trained using more data from some countries than from others; these biases need to be studied.

It would be necessary to perform this study considering other (language) versions of Wikipedia in order to understand how an editor's language affects the coverage bias.

Reproducibility
Code implementing our methods, the merged list of floods, and the raw and processed datasets of Wikipedia matches are available: `https://github.com/javirandor/disasters-wikipedia-floods`.

In this chapter we noted how user-generated data can carry biases. As we expect the same from other social networks, we can't draw conclusions on the severity of a phenomenon based 'solely' on the intensity of its activity on social media. The next chapter confirms that this is also the case for Twitter.

Part II

Social Media Analysis for Situational Awareness

Chapter 4

SOCIAL MEDIA AS A SOURCE OF FLOOD-RELATED INFORMATION

In the current chapter, we describe how to integrate social media into EFAS to provide valuable signals augmenting flood risk information provided by this platform, by finding potentially relevant and representative messages from flood-affected areas in the languages spoken in those areas.

4.1. Introduction

Although the value of social media analysis in providing timely data and methods for the analysis of natural hazards has been recognized in previous work, comparatively much less attention has been given to how to integrate social media in a seamless, reliable way with tools for disaster forecasting and monitoring.

Since the geographical domain of EFAS products covers an area where population speaks more than 27 languages, we focused on a multilingual system. We therefore use representation of words as vectors in order to exploit probabilistic functions to infer similarities between words, know as word embeddings[73].

Our research fills this gap by describing the integration of social media monitoring into a flood monitoring and forecasting platform, enriching hydro-meteorological

information with reports from the public.

We developed software for EFAS named Social Media for Flood Risk (SMFR) which provides near-real-time information collected from social media about flood risks and impacts, including examples of messages in social media about it. Figure 4.1 represents the conceptual schema of SMFR's components, part of which is described in this chapter, and their integration.

Our main contributions are:

- We integrate social media data collection into EFAS based on its forecasts. Whenever EFAS rapid risk assessment identifies heightened risk of floods in a certain area, we trigger data collection from social media (in our case, Twitter) respecting API limitations while dealing with the possibility of various events happening at the same time. A similar mechanism is already in use for triggering pre-tasking of satellite image acquisition in Copernicus EMS

- We describe a methodology that requires a minimal amount of manual intervention for each additional language, and demonstrate it with four languages being used to bootstrap a classifier for a fifth language. This methodology is based on Convolutional Neural Networks and its multilingual capabilities stem from using either language-agnostic word embeddings, where vectors representing sentences are not dependent on a single language, or multilingual word embeddings [13].

- We describe an aggregation and selection module that can select representative messages for an area in which flood risk has been predicted.

- We integrate SMFR into EFAS and demonstrate it during the recent floods affecting Calabria, Italy, in early October 2018. Note that we decided to present only one case for the sake of clarity and conciseness, however SMFR has been tested with additional real cases.

The goal of our system is twofold: while bringing more detail to the outcomes of an hydrological model seems to answer to our *RQ1: Is it possible to integrate effectively social media signals with authoritative data at a pan-european level where riverine flood likelihood is estimated?* expressed in Chapter 1, Our methodology offers an answer also to the second question *RQ2: Is it possible to classify reliably the relevance of social media information to floods using a 'zero-shot' transfer learning ?*.

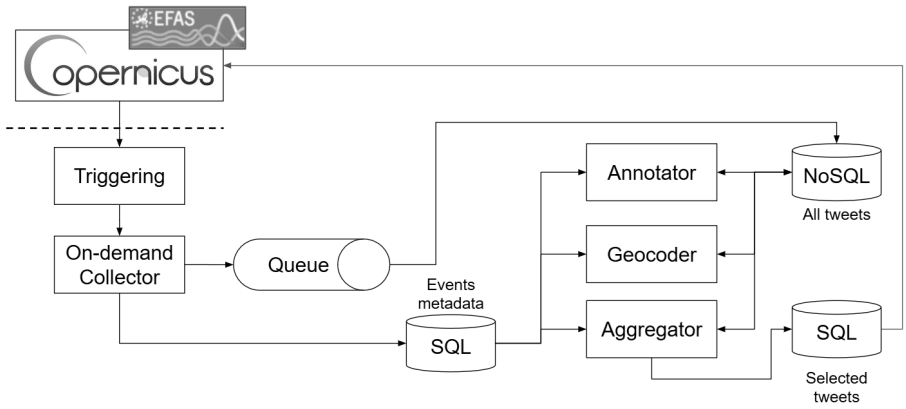


Figure 4.1: Schema of SMFR components

The remainder of this chapter is organized as follows: the next section presents related work; then, three technical sections describe the methods for on-demand data collection, multilingual classification, and aggregation and selection. Finally, we present experimental results, including a case study, followed by our conclusions and future work.

4.2. Related Work

In previous work, authoritative and nonauthoritative data have been combined in various ways.

Multilingual classification of social media postings
SMFR is designed to work across multiple languages. The main processing that we do to messages is to determine whether they are relevant to flood risks/impacts or not. This is done through supervised classification, which requires labeled data. However, to work across multiple languages in practice requires to be able to classify messages in languages for which we may not have labeled data yet.

Past research addressed cross-lingual bootstrapping of classifiers for natural disasters detection on twitter [63] relying on automatic translation to use available models. Khare *et al.* (2018) built a statistical-semantic classification model

with semantics extracted from BabelNet and DBpedia and compared relevancy classifiers with datasets translated into a single language, as well as with cross-lingual datasets. It was shown how adding semantics increases cross-lingual classification accuracy. increases cross-lingual classification accuracy.

Our work, in contrast, does not require semantic resources, we only leverage on word embeddings for multilingual modelling as demonstrated by Luong *et al.* Indeed, we present and demonstrate two different ways in which word embeddings can be used to perform multilingual classification: using language-agnostic word embeddings learnt from a multilingual corpus [73], and using multilingual word embeddings that are aligned across languages [13]. In both cases, we can use labeled data in a set of known languages to bootstrap a classifier for a new language for which no labels are available.

On-Demand data collection

The data collection should ideally achieve high recall, capturing a large fraction of the relevant information, while at the same time having high precision, avoiding irrelevant information. Both goals usually enter into conflict and trade-offs are necessary. Previous research has described extensively how data collection from Twitter is done [35]), in this section we focus on the specific aspects of our system, which performs *on-demand data collection*.

The key element of our data collection is its triggering mechanism which is done dynamically according to flood forecasts. EFAS runs two simulations per day, identifying NUTS-2 areas (typically regions or provinces) where there is a high probability of floods impacts in the following 48 hours as shown in figure 4.2. Once the list of NUTS-2 areas is received, the system extracts their coordinates and the names of all cities in the area that have more than a certain number of inhabitants. The coordinates are used for filtering the Tweets by location while the names of cities, in english and local language, translate into a series of 'OR' filters by keywords. In our current configuration, we set this threshold empirically to 80,000 inhabitants, which tends to capture a handful of cities for each event. A lower threshold is possible as long as it does not generate a large number of city names that exceeds Twitter's API limitations.

Given the limitations in Twitter's API, we use a single connection against Twitter's public streamer at any given time. Additionally, the public streamer limits queries to up to 400 keywords (each of less than 60 bytes) and up to 25 location rectangles.[1] A *query builder* component groups several active events

[1] `https://developer.twitter.com/en/docs/tweets/filter-realtime/`

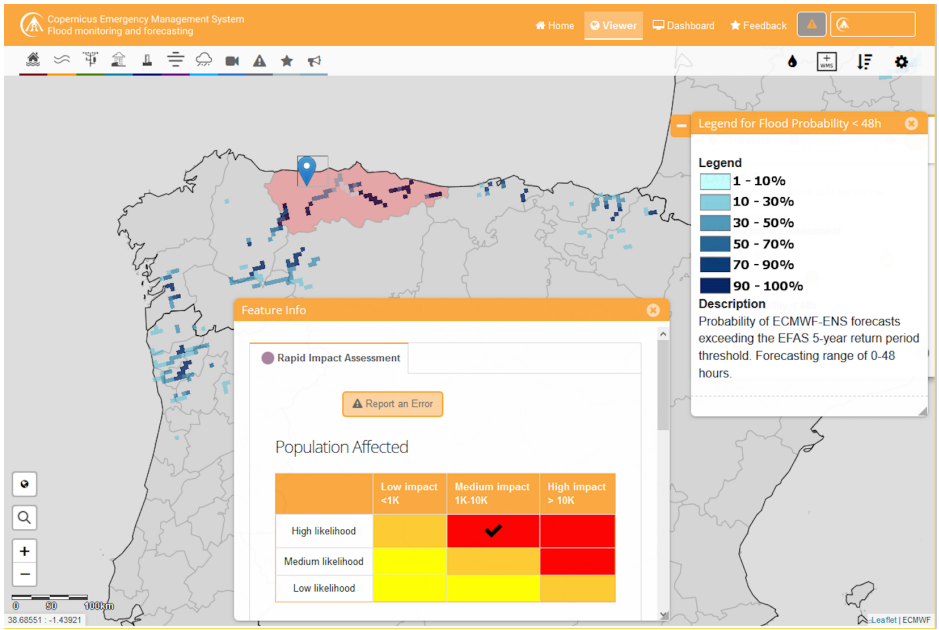


Figure 4.2: Screenshot of EFAS web interface with the layers identifying areas where there is high probability of floods in the following 48 hours and rapid impact assessment

in order to update a single request respecting these limitations, while keeping information about the single possible event identified by EFAS. Each collection is kept active for two days after the expected peak time. If the peak time estimation is updated by a new EFAS simulation, the collection's expiration time is extended. Given that we have a single query, SMFR has to separate the incoming stream into different events according to locations and keywords. If a message belongs to overlapping regions or contains names of cities in different events, the message is copied to all the matching events.

Figure 4.3 depicts how rapid risk assessment leads to the definition of a series of keywords (city names) and locations (rectangles containing NUTS-2 areas) for filtering information from Twitter's public streamer. The areas in yellow and red in the figure are identified as having high risk of flood by EFAS. Then, the

`overview` accessed November 2018.

system determines names of cities in each area and bounding boxes surrounding each area. We have heuristics that merge neighboring areas and de-prioritize smaller cities if their quantity exceeds Twitter's limitations, but in practice these are rarely triggered.

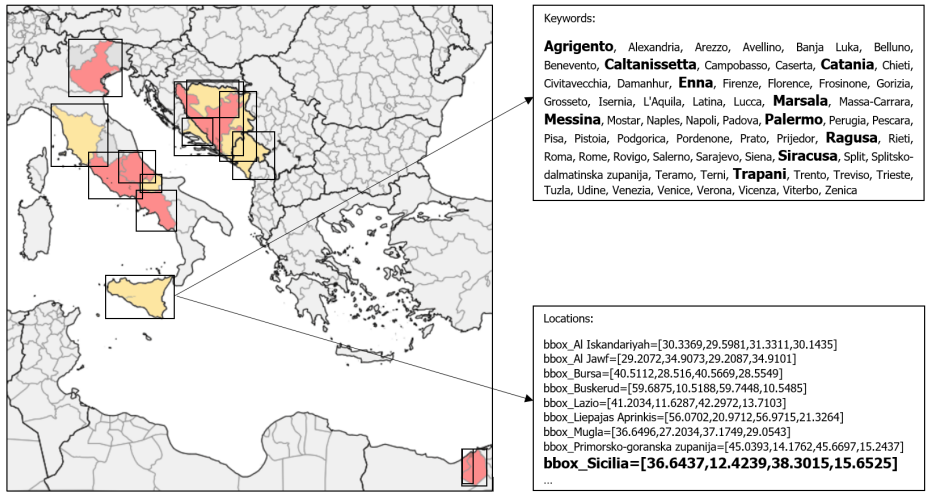


Figure 4.3: Example depicting areas in which EFAS forecasts high risk, which appear in yellow and red color. Each area defines a set of keywords, which are names of cities, and locations, which are bounding boxes (rectangles). These keywords and locations are used to gather information from Twitter.

Multilingual classification

The setting used was supervised binary classification. The positive class comprised all messages indicating that "a flood has just happened or is about to happen" while the negative class included all other messages.

Training data was labeled using crowdsourcing platform GetHybrid (`https://gethybrid.io/`). The amount of labeled data comprises over 7,000 annotated messages, each one annotated by three annotators independently. Four training datasets were created, containing labeled tweets in German, English, Spanish, and French. Each language includes between 1,200 and 2,300 annotated messages, as shown in Table 4.1 (column "TL" for Total Labeled) in the section describing the experiments. This amount of data is typical in automatic

classification tasks [35].

Convolutional neural networks

We tested a number of learning algorithms, including Support Vector Machines [39], which have been shown to be effective for a number of general text classification tasks, and Random Forests, which have been shown to be effective specifically for classifying crisis-related Twitter messages [36]. In both cases, we represented messages as bag of word unigrams and bigrams, and the performance was better than simpler methods such as a naïve Bayes classifier or a decision tree.

A comparatively newer approach for text classification that has proven to be quite effective is the use of Convolutional Neural Networks (CNN). These have been employed for a number of tasks including sentiment analysis [91].

There are four main operations in every Convolutional Neural Network:(i) Convolution; (ii) Non Linearity, Rectified Linear Unit (ReLU); (iii) Pooling or Sub Sampling;(iv) Classification (Fully Connected Layer). The primary purpose of (i) Convolution is to extract features from the input using a filter smaller in size than the original input. The second (ii) operation ReLU replaces all negative values in the feature map by zero. The (iii) Pooling operation , max pooling in our case, reduces dimensionality by taking the largest element from the rectified feature map within a neighborhood. The (iv) Classification operations, after a series of iteration of the (ii) and (iii) then uses the higher-level features identified to deter a class using several fully connected perceptrons layers.

A Support Vector Machine (SVM) is a classifier defined by a separation hyperplane. Input of the algorithm is the labeled training data, while its output is an optimal hyperplane which places new data into distinct classes.

Random forest classifiers infer a series of decision trees from randomly selected subset of the labeled data. The results are then passed through different decision trees to finally classify the test data.

The specific architecture we used is described in the next section; using it, we performed various experiments using 10-20 epochs of training (passes over the entire training set in randomised order).

The results were similar in performance to the SVM and Random Forests. However, manual error analysis showed that qualitatively these errors were different. While in the case of SVM and Random Forests a misclassification, such as a false positive, is usually the result of a word marking flood relevance (e.g.,

"flood") used in a completely different context (e.g., "my timeline is flooded with messages"), in the case of the neural network we used a misclassification was often semantically related to floods, such as a message referring to the effects of other natural disasters. The lesson learned from these experiments was that neural networks are better in this problem at capturing semantic characteristics that are relevant for our task of distinguishing flood-related messages.

Neural network architecture

Current implementations of convolutional neural networks for text processing tasks tend to have a similar architecture. They consist of an input layer, a word embedding layer, a series of convolutional and max pooling layers, a dense layer, and an output layer [47].

The input layer holds a padded sequence of words with a maximum length $S = 100$ words, which is more than sufficient for tweets considering their maximum length is 280 characters.

The word embedding layer converts every word into a low-dimensional vector, typically in the order of a few hundred dimensions (e.g., $D = 200$ or $D = 300$). We used two sources of pre-trained word embeddings, as described in the next section. For each pre-trained word embedding, we considered two configurations-ups: one in which the parameters of the pre-trained word embedding were fixed, i.e., not modifiable while training the neural network, and one in which they were part of the optimization process, i.e., modifiable while training. In our experiments, best results were obtained when these parameters were fixed, probably because the amount of flood-specific data that we are using for training is small in comparison with the corpora used to create these word embeddings. The results we report on this chapter use fixed word embeddings.

The convolution layers collect several word embeddings representing adjacent words and "summarize" them into a single vector. The main parameter for the convolutions is the width C, which is how many adjacent words to take into account. In the text "flood warning due to heavy rain" using $C = 5$, there are two possible convolutions: "flood warning due to heavy" and "warning due to heavy rain." The parameter C is determined considering what is the effect of the context on the meaning of a word, and $C = 5$ is a typical value. We did not observe any increase in performance with a larger value of C, while a smaller value of C may lose contextual information.

The max pooling layer collect a series of m disjoint convolutions as input,

and generate a vector of dimension $d < D$ as output. The max pooling step operates differently from the convolution layers in the sense that the windows it uses are disjoint, i.e., non-overlapping. The purpose of this layer is to reduce the dimensionality of the network for computational purposes and to reduce the chances of overfitting. In our case we used $m = 5$ and $d = 128$, which are typical parameters used in text classification.

The final layer is a densely-connected (complete) layer. All the neurons in the last max pooling layer are connected to all the neurons in this dense layer, and all the neurons in this dense layer are connected to the two output neurons. One of the two output neurons should activate when the example is positive (i.e., a message indicating that a flood has just happened or is about to happen), and the other output neuron should activate when the example is negative (i.e., the message does not indicate that a flood has just happened or is about to happen).

Word embeddings for ultilingual classification

The usage of word embeddings allows to incorporate multilingual capabilities in two ways: by using language-agnostic word embeddings and by using language-aligned word embeddings.

Our source of language-agnostic word embeddings is GloVe [73], which are vectors of dimensionality 200 obtained from a large corpus of tweets containing 27×10^9 tokens (1.2 million of them unique). While these word embeddings were not developed for multilingual tasks, they do incorporate any word present on a tweet in a language-agnostic manner.

Our source of language-aligned word embeddings is MUSE [13], which are vectors of dimensionality 300 obtained from various snapshots of Wikipedia in various languages. For each language, vectors for the 200,000 most frequent tokens are provided, and these vectors have been *aligned* across languages using parallel lists of tens of thousands of words. In the resulting embeddings, two words with the same meaning in different languages are mapped to similar vectors.

In our experiments, presented on the next section, both pre-trained sets of vectors allow to transfer an automatic classifier learnt with labeled data from one language (or a set of languages) into another language with no new labeled data ("cold start") or with a limited amount of labeled data ("warm start").

4.3. Methods

In our integration with EFAS, flood risk is established by an hydro-meteorological model and our task is to add complementary information that brings a better understanding of the situation on the ground. Relevant messages are mapped to NUTS-2 areas (Nomenclature of Territorial Units for Statistics, Level 2) either by using explicit coordinates which are rarely present in tweets, or more often, via a text-based geocoder. Geocoding deals with messages that do not include explicit geographical coordinates, but mention a place name such as landmark or city. Geocoding uses a Named Entity Recognition tagger to obtain possible locations in a text considering the syntax of the message. It uses a gazetteer in a large database of place names with their corresponding geographical coordinates, and finally it uses a neural networks to infer the correct country and correct gazetteer entry for those places. We used a library named Mordecai [29] which extracts place names from a piece of text, resolve them to the correct place, and return their coordinates and structured geographic information.

Messages are aggregated at the level of an event, but also at the level of each NUTS-2 area. In both cases, one key operation is to select a representative subset of messages. We do this operation by using the following efficient heuristic based on de-duplication and text centrality:

1. Select up to 5,000 tweets having at least a 90% probability of being flood-related; if there are more than 5,000 tweets, select the ones with the highest probability

2. Compute similarities between these tweets

   a) Consider only pairs having probabilities of being flood-related that differ at most by 0.0001, exploiting the fact that near-duplicate tweets will be given the same probability by the neural network

   b) Compute edit distance and use it to compute normalized similarity: $1 - \ell(m_1, m_2)/(|m_1| + |m_2|)$ where $\ell(m_1, m_2)$ is the edit distance between the two messages and $|m_1| + |m_2|$ the sum of their lengths.

   c) If the normalized similarity is greater than 0.8, use the timestamps of the tweets to mark the newer tweet as a duplicate of the older tweet

3. Sort all unique tweets by their *multiplicity*, i.e., by the number of duplicates they have, and keep the top 100

4. For these tweets, compute all pair-wise similarities using the same formula as above, and add the similarities for every tweet; this is the *centrality* of the tweet.

Step 1 of the heuristic has the goal of removing messages that are irrelevant. Step 2 removes near-duplicates which are redundant, but keeps track of how many near-duplicates a tweet has, for the purposes of using redundancy as a signal of importance. Step 3 applies a well-known lexical centrality heuristic [21], in which a salient message is one that has content in common with many other messages.

Figure 4.4 shows a first deployment of SMFR on EFAS web interface, it describes how the areas and their most relevant tweets are presented to EFAS users. The NUTS-2 area can be grey (low activity), orange (medium activity) or red (high activity) according to the ratio between numbers of annotated tweets. The triggering that lead to the creation of the collection depicted can be seen in Figure 4.2

Experimental results

We describe two types of experimental results. First, we perform experiments to test the performance of the multilingual-classifier, comparing it with a monolingual classifier. Second, we show a real example of an actual flood event, describing the performance of the on-demand collector, of a multilingual classifier that does not use labeled data from the target language, and of the aggregation and selection method.

Multilingual classifier

The multilingual classifier provides a solution for bootstrapping classifiers in new languages based on labeled data for other languages. The aim of the experiments is to compare the classifiers trained with and without labeled data for a specific target language. Performing this evaluation, nevertheless, requires having labeled data to measure effectiveness parameters.

For these experiments we use four labeled sets of flood-related tweets in German (DE), English (EN), Spanish (ES) and French (FR). These were labeled by crowdsourcing workers via crowdsourcing as previously described; the question the annotators had to answer was "is this message indicating that a flood is happening or about to happen?" Hence, positive examples are the ones that are related to flood risk and impact, and negative examples are the ones that are not.

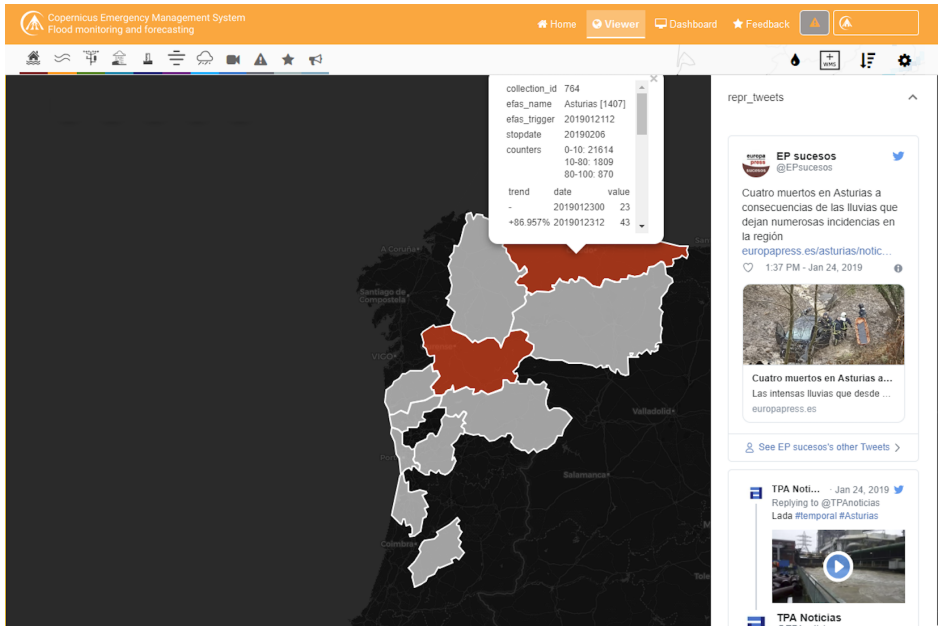


Figure 4.4: Screenshot of EFAS web interface with the layer identifying areas where there is high tweet activity and their most representative tweets on the right side. The basemap has been darkened for better visualization

For each language and word-embedding source we performed three experiments: monolingual, cold-start, and warm-start. We used two thirds of the labeled data for training the classifiers and the remaining third for testing. We keep the testing portion fixed across experiments.

In the *monolingual* experiment we simply use labeled data in one language to predict the label for messages in the same language. In the *cold-start* experiment we train a classifier for a new language using only labeled data for other languages; for instance for automatically labeling tweets in Spanish, we use a classifier trained on labeled data for English, German, and French. In the *warm-start* experiment we use a set-up similar to the one of the cold-start experiment, but we add a limited number of messages (300) labeled in the target language. Each experiment is done once using GloVe embeddings and once using MUSE embeddings. We report precision, recall, and F-measure for each experimental setup in Table 4.1.

Table 4.1: Classification results for four languages (German, English, Spanish, and French). TL indicates the total number of labeled messages, while Pos. indicates the percentage of those who were labeled as flood-related. P, R, and F indicate Precision, Recall, and F-Measure respectively. We report the performance of a monolingual classifier, of a cross-language classifier with "cold start" (uses no training data in the target language), and of a cross-language classifier with "warm start" (uses 300 labeled items in the target language).

			Glove embeddings									MUSE embeddings								
			monolingual			cold-start			warm-start			monolingual			cold-start			warm-start		
	TL	Pos.	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
DE	2356	46%	0.95	0.82	0.87	0.59	0.85	0.70	0.93	0.8	0.86	0.88	0.85	0.87	0.54	0.82	0.65	0.89	0.80	0.84
EN	1999	20%	0.79	0.63	0.70	0.59	0.49	0.54	0.67	0.51	0.58	0.64	0.68	0.66	0.33	0.50	0.40	0.58	0.28	0.38
ES	1592	48%	0.80	0.78	0.79	0.61	0.75	0.67	0.71	0.83	0.77	0.70	0.84	0.77	0.62	0.69	0.65	0.68	0.89	0.77
FR	1248	40%	0.74	0.72	0.73	0.50	0.46	0.48	0.62	0.77	0.69	0.69	0.75	0.72	0.44	0.86	0.58	0.59	0.72	0.65

Table 4.1 reports values for Precision (P) , Recall (R) and F-measure. We also include the total number of tweets manually labeled by crowdsourcing workers (TL) and its percentage of positive (Pos.) tweets. Collections for German, English, Spanish, and French consider tweets posted during floods happening in the last two years in Germany, the UK, Spain and Mexico, and France respectively. Hence, the number of labeled tweets and the percentage of positive examples differ across languages.

All the results show the same pattern: the monolingual classifier performs best, as expected; the cold-start classifier (which does not use any labeled data in the target language) suffers from a loss mostly of precision, but also of recall; and the warm-start classifier (which involves annotating a small number of tweets in the target language) has better performance than the cold-start classifier both in terms of precision and recall. Indeed, the warm-start classifier often achieves an F-measure that is comparable to the one of the monolingual classifier. Regarding the choice of word embeddings, results suggest that the performance using GloVe or MUSE embeddings are comparable.

In general, considering the combination of the information from the classification with the known locations from the EFAS forecasts, the classification performance is sufficient to extract representative tweets from an event and to map approximately the affected locations, as we demonstrate next.

4.4. Case Study: Calabria Floods in October 2018

In early October 2018, floods affected the region of Calabria in southern Italy. At least 2 people died in flash flooding after severe weather which peaked on October 5th. A mother and her seven year old son, who were swept away by flood waters in their car, were found in a river near Lamezia Terme, between the towns of San Pietro a Maida and San Pietro Lametino in Calabria. Other areas of Calabria were also hit by flooding and landslides.

Several families were forced to evacuate their homes and people were rescued after they climbed onto the rooftops of houses to escape the flooding.[2] Italian news agency ANSA, stated that the Ponte delle Grazie bridge on provincial highway 19 in the area collapsed during the storms [78]. Vigili del Fuoco, Italy's National Firefighters Corps, reported major flooding in Ciro Marina, Petilia de Policastro, Strongoli, Cotronei and Isola di Capo Rizzuto. As shown in Figure 4.5 (a) more than 300 mm of rain fell in 3 days [90].

EFAS forecasted a potential flood in the Calabria NUTS-2 area on the 4th of October with a predicted peak time of the event for the following day. As planned, SMFR triggered a collection with a duration of 2 days that was later extended for an additional day due to persistence of the signal from EFAS forecasts. We analyzed the collection once it was stopped, at midnight on the 7th of October, after collecting 14,347 tweets.

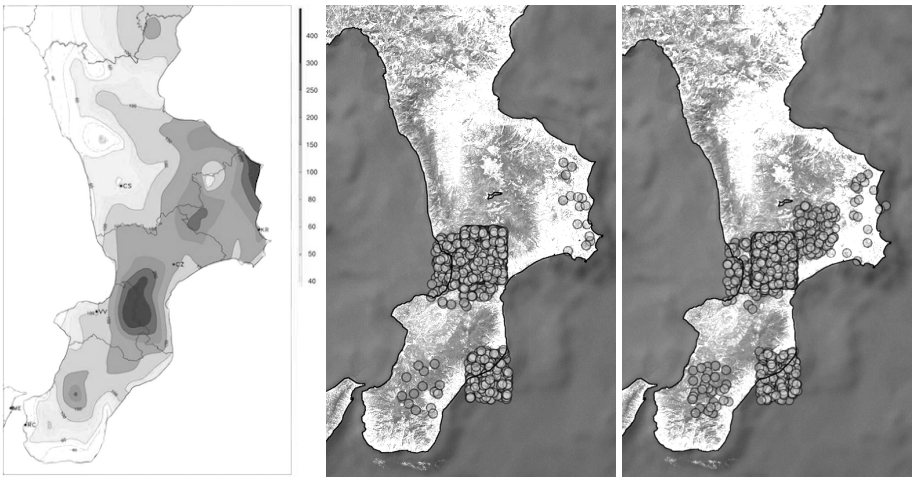
In order to confirm what emerged from experiments in the previous section, we trained two classifiers for messages in Italian, the first (cold-start) using only labeled data in German, English, Spanish, and French, and the second (warm-start) adding 300 manually labeled tweets in Italian from the collected dataset. For brevity we present results obtained using the GloVe embeddings (results using MUSE embeddings are similar).

In Figures 4.5 (b) and (c) we depict the position of geo-located tweets annotated by the cold-start and warm-start classifier respectively. Tweets have been filtered using a relevance to flood (label predicted) greater or equal to 0.8. We include the tweets geolocated within the bounding box used for triggering the collection, resulting in 2,847 tweets for the cold-start scenario against 3,857 for the warm-start scenario. For visualization purposes, tweets geo-located to the exact same location are randomly scattered by a small amount in the map.

Figure 4.5 confirms the results from the experiments, in the sense that both

[2]Data provided by `https://floodlist.com`, which is a EU-supported project providing reports on past floods.

cold-start and warm-start classifiers are able to classify relevant tweets, with an advantage for the warm-start classifier in the sense that it identifies more relevant tweets and has better coverage of the areas affected by heavy rainfall. This suggests that the cold-start method can provide a first approximation for identifying an ongoing event, while the warm-start method yields more precise and relevant tweets.



(a) Rainfall **(b) Tweets, cold start** **(c) Tweets, warm start**

Figure 4.5: Comparison of (a) rainfall, (b) tweets located by cold-start model, and (c) tweets located by warm-start model. Data from floods in Calabria, Italy, 2-5 October 2018. Tweets falling in the same location are randomly scattered for visualization purposes.

Finally, Table 4.2 shows samples of tweets that have been selected as the most representative for this event, following the heuristic described previously, and taking as input the tweets found relevant using the cold-start and warm-start methods. We can see that in the cold-start scenario the most representative tweets are relevant to the event considered. However, the warm-start scenario gives more informative messages. While tweets selected by the cold-start classifier are relevant, they mostly just confirm the event as reported in the news; in constrast, the tweets selected in the warm-start case identify a message from the Italian Prime minister as the most representative ("I follow with concern the evolution of events ...") and include information about damages and casualties due to the

flood.

Table 4.2: Representative tweets selected by cold-start and warm-start. Conf. is the confidence of the classifier. Mult. the multiplicity (number of near-duplicates of the tweet). Cent. is the centrality (number of closely related but not duplicate tweets).

Cold-start			
Conf.	Mult.	Cent.	Text (first ∼10 words)
1.0	87	89	Second flood in Calabria in 40 days. Devastation and 2 casualties ... *(Seconda inondazione in Calabria in soli 40 giorni. Devastazione e 2 vittime ...)*
1.0	11	93	Bad weather in Calabria, the kennel is flooded ... *(Maltempo in Calabria, il canile e 'sommerso dall' acqua ...)*
1.0	7	94	Bad weather: Red alert in Calabria today and in Puglia tomorrow ... *(Maltempo: oggi allerta rossa in Calabria e domani in Puglia ...)*
1.0	5	97	Meteo, panic in Calabria: streams flooding roads. Rescuers using rubber boats ... *(Meteo, caos in Calabria: torrenti esondati e strade allagate. Soccorsi in gommone ...)*
1.0	5	87	Bad weather in Calabria, missing mother and her two sons found dead ... *(Maltempo Calabria, trovati morti mamma e due bimbi dispersi ...)*
Warm-start			
Conf.	Mult.	Cent.	Text (first ∼10 words)
1.0	194	76	I follow with concern the evolution of events in #Calabria ... *(Seguo con apprensione l ' evolversi degli eventi in #Calabria ...)*
1.0	194	88	Water bomb in Calabria, among the upset in the population ... *(Bomba d ' acqua in Calabria, tra la popolazione sconvolta ...)*
1.0	14	46	# breakingnews Bad weather Calabria: a woman and one of her son found dead. ... *(#ultimora Maltempo Calabria: morta una donna e suo figlio, disperso il fratello ...)*
1.0	23	98	Bad weather in Calabria, mom and son found dead, missing 2yrs old brother ... *(Maltempo in Calabria, morti mamma e figlio: si cerca il fratellino di 2 anni ...)*
1.0	8	94	Bad weather, nigthmarish night in Calabria, Civil Protection: "High risk" ... *(Maltempo, notte da incubo in Calabria, Protezione civile: "rischio vittime" ...)*

4.5. Conclusions

Our work provides a solution and methodology for integrating flood modeling and evidence from the ground in real-time for several countries, potentially providing information from local witnesses or local media to first responders. This unique combination of hydrological simulation forecasting and an automatic, immediate monitoring of the extent of the event through social media without necessity to manually translating information, allows to shorten the response time, which is extremely precious in the very early stages of a flood. Moreover, during the development of an event, collected messages could be

valuable to international rescue coordinators such as ERCC because they provide insights about the local response, about whether alerts that have been issued by authorities, and about some of the concerns that those affected by a flood or a flood alert may have. The research also highlighted the need for high-recall data collection in which data in multiple languages is captured, and provided a methodology for dealing with a new language, by bootstrapping a classifier with similar languages for which labeled data is available, using either language-agnostic or language-aligned word embeddings. Additionally, it was clear during the development of the project that naturally occurring data (i.e., actual messages posted during a flood in a particular country) are necessary to build an accurate classifier and aggregator.

Limitations
Due to the nature of EFAS and its geographical domain (pan-European) we focused on language spoken in the region, therefore we only tested the methodology proposed for indoeuropean languages. The performance with other languages remains to be studied.

At the time of writing, the described system is still in testing phase. After a period of internal evaluation, SMFR will become operational and its results will be disseminated among EFAS Partners.

We can envision a global system comprising dozens of languages used to augment GloFAS coverage. One can also envision further steps in the direction of using social media as a data source that can feed into a predictive model, using it not only for confirming known flood risks, but also for detecting new ones, particularly in areas where digital devices are prevalent but meteorological stations and other physical sensors are scarce in comparison.

Reproducibility
Datasets and code for the experiments described on this chapter will be available for research purposes at `https://bitbucket.org/lorinivalerio/iscram\_2019/`.

After having described how our research can help detecting floods, our work continue to study the possibility of assessing impacts during the development of a flood.

Chapter 5

SOCIAL MEDIA AS AN ALERT SYSTEM

Given the importance of making well-informed decisions, our research proceeded to answer our *RQ3* (i.e., the possibility to detect floods worldwide from social media reports). We run experiments to indicate that a model can indeed spot impactful events where damages are clearly related to water.

5.1. Introduction

Even if the Paris Climate Agreement, which came into force in 2016, succeeds in keeping the global average temperature rise well below 2°C compared to pre-industrial levels, 'global warming' is still expected to cause severe impacts. Under the most optimistic scenario of a 1.5°C warming, flood damage is nevertheless set to increase by between 160% and 240% [19]. Regional crisis management organizations in wealthy countries can afford the cost of high-resolution flood-monitoring systems. On the other hand, international relief organizations with a global scope rely on global hydrometeorological models. Different socioeconomic realities combined with heterogeneous data availability (the so-called 'data divide' [27]) translate into various degrees of uncertainty, with reliable flood forecasts often possible only for big events.

Figure 5.1 shows maps taken from GloFAS, illustrating the probability of the daily streamflow forecast to exceed the local '1 in 20-year' discharge (i.e., the 20-year threshold, considered a severe event). GloFAS works by running multiple perturbed simulations, and the probability of a peak discharge exceeding the 20-

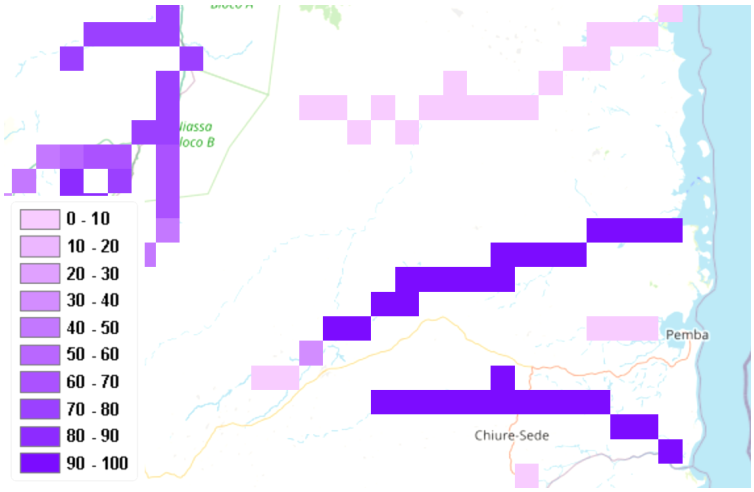


Figure 5.1: Examples of low-uncertainty (Megaruma River in Mozambique) streamflow forecasts

year threshold is the fraction of such simulations above this threshold. Larger values, or darker color in Figure 5.1, indicate a reduced uncertainty, as most simulations agree on forecasting a severe flood event for the day.

For instance, in the low-uncertainty forecast used as an example in Figure 5.1 (top), for a few river branches almost all the simulations converge (100% probability)

The uncertainty is reduced when the flow peak is predicted to occur within few days when the forecast is mostly driven by hydrological rather than meteorological conditions, and it is therefore more reliable. High uncertainty is often associated with a large lead-time of the prediction, and the absence of a clear flood signal in meteorological forecasts.

Previous research has sought to integrate social media information into flood monitoring systems. This research, at the intersection of crisis informatics and disaster risk reduction, has been based largely on the extraction of public-generated discussions about flood risk in situations where weather alerts have been issued by relevant authorities, and reporting of the concerns of those impacted [51, 14]. Such systems are affected by the same limitations and uncertainties as the hydrometeorological forecasts themselves.

Against this background, the central question addressed by our research was:

'Is it possible to identify floods worldwide from social media reports, using knowledge from past events and independently from hydrological forecasts?' expressed in Chapter 1 as *RQ3*. Using machine learning, we created a model that takes as input the volume, trends, and characteristics of discussions about floods in social media. The output of our model is the probability that an actual flood happens, computed by supervised learning based on past events. Because the data source of social media which we use (i.e. *Twitter*), despite its large coverage and volume, produces a noisy signal that does not yield high-accuracy alerts, we cannot positively and conclusively answer the posed research question. However, our work suggests that the question may be partially answered in the affirmative, in that we can complement a flood forecasting system reducing the uncertainty of hydrological forecasts. Social Media information could be seen in this case as additional support for the Crisis Managers in the decision-making process.

In the following sections, an overview of related work is first provided, and the methods for creating a training dataset and building the model for event detection are described. Finally, the experimental results of our work are presented, followed by conclusions and priorities for future work.

5.2. Related Work

This section provides an overview of some flood detection systems based on hydrometeorological models, social media or both.

Flood detection with hydrometeorological information

NASA's real-time Global Flood Monitoring System (GFMS) is driven by precipitation information from the joint NASA - Japan Aerospace Exploration Agency (JAXA) satellite missions - the Tropical Rainfall Measuring Mission (TRMM) and its successor, the Global Precipitation Measurement (GPM) mission [108]. GFMS performs rainfall analysis using a physically based hydrological model, and has a detection performance that is highest for floods of long duration and affecting a large area.

The previously mentioned GloFAS, developed jointly by the European Commission and the European Centre for Medium-Range Weather Forecasts (ECMWF), is a global hydrological forecast and monitoring system independent of administrative and political boundaries, that is fully operational within the EU's Copernicus Emergency Management Service. GloFAS couples weather forecasts with

a hydrological model to produce daily flood forecasts. Due to its meteorological forcing (i.e., rainfall map, wind speed map, temperature map, etc.) and spatial resolution of 0.1 degrees, GloFAS performs well for large rivers. The lack of finely distributed meteorological observations at a global scale limit the resolution of the calibration for the forecasting for smaller rivers [32].

Flood detection from social media
Geospatial information relevant to an event are usually available after the events have ended, and are generally provided by satellites images or field surveys. Recent studies have focused on whether combining social media and geo-information can speed up early warnings [105].

Geospatial information relevant to an event are usually available after the events have ended and provided by satellites images or field surveys. In contrast with our research, we analyze how information derived from social media can be used independently from physically based models in detecting several flood-types (e.g., riverine, flash floods, hurricane floods). The novelty of our approach is that, unlike the works mentioned above, it does not rely on any leading factor and does not restrict the nature of events analyzed. We use a validated reference set of events tracked by independent organizations, covering a wide range of events in terms of geographical region, duration, extent and magnitude.

Based on this we have built a supervised model for catching signals from social media on heterogeneous types of flood (riverine floods, coastal floods, flash floods, hurricane floods, etc.) at a sub-national scale. We then analyzed our results in the light of forecast information available at the time, just before the event, in order to understand if social media information can represent an added value to those systems.

5.3. Methods

The ground truth data that we used is constructed at the level of days and countries, with each record indicating if there was a confirmed flood on that day in that country. We collected social media data and then processed it to the same level of granularity by extracting various features. Further details on the dataset preparation are presented below.

Ground truth
Because no single comprehensive database exists that containing all worldwide

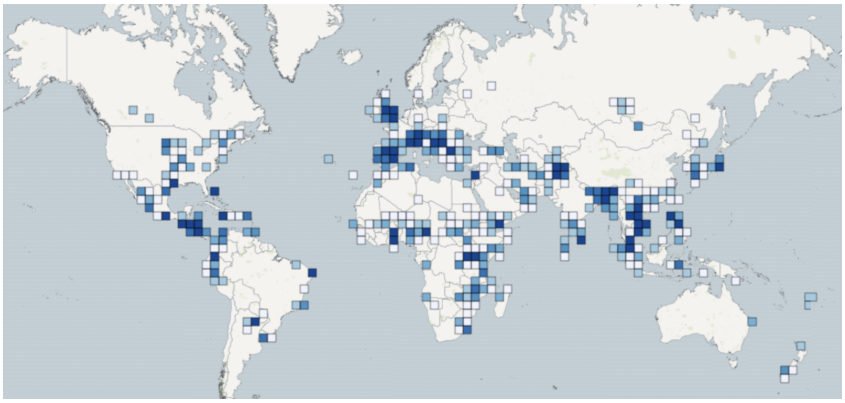


Figure 5.2: Spatial distribution of the flood events used as ground truth. Darker color indicates multiple events in the same administrative area

floods, we used a list of flood events from previous research that aggregates data from different sources. For details on this list of events, the interested reader can consult the original source [52]. Briefly, the list of flood events is collected from three different databases: Europe's Floodlist; the UN's Emergency Events Database (EM-DAT), and the Dartmouth Flood Observatory (DFO) of the University of California.

We selected all events that could be geocoded to a sub-national level. This led to 349 events spread over 1,318 administrative areas, as shown in Figure 5.2.

Data collection

We collected tweets relying on the public Twitter streamer.[1] We opted to collect posts on floods using a set of flood-related keywords in several languages (i.e., English, Spanish, Italian, German, French, Portuguese, Arabic) for a nine-month period, covering flood-seasons worldwide. The complete list of keywords is available in our data release. Our data collection period was April to December 2019, but was interrupted frequently. We experienced network, software, and hardware failures, together with limitations applied by the streamer provider. We worked around the multiple interruptions during the 274 days of observation, by applying the following heuristic method: if for any given day there were more

[1] `https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/get_statuses_sample`

than six hours without any tweet, we would mark that entire day as 'invalid'. In our observations, this was always an indication of some kind of failure. The days for which we have data were 74% of the total observation days.

Features

Each collected tweet was automatically annotated as either flood-related or not flood-related by a multilingual classifier [51]. The flood-related tweets were then geocoded using their available geographical metadata, and when this was not present (i.e. in the majority of cases), using place-names mentioned in the text[29]. Next, we aggregated individual tweets in space (i.e. the affected area) using the Database of Global Administrative Areas (GADM) spatial database of the location of the world's administrative areas.[2]

We aggregated tweets according to days, to match the granularity of our ground truth data.

For each day and region, we measured the features listed in Table 5.1.

In order to enable comparisons between regions with different population sizes and different degrees of Twitter adoption, we produced normalized features by considering the average number of postings originating from each region during a period of one month. We calculated such values by analyzing one month of geolocated data extracted from the Public Streamer available on the Internet Archive digital library.[3] All of the 72 features in Table 5.1 were divided by the expected number of postings for the same region, with the exception of features P00-10 ... P90-100 and T3P00-10 ... T3P90-100, which reflect proportions, and the 'language' feature. In addition, the expected number of postings for the region was added to the feature list, for a total of 73 features.

Data labeling

In order to create the training data, it is not sufficient merely to associate each row in the dataset table (corresponding to a date and a region), with a label of flood (*True*) or no-flood (*False*). The main reasons for this are that flood events and the related discussion on social media generally last more than one day, and they build up over time. Furthermore, there is no common, widely accepted definition of when a flood starts or ends. Indeed, in the original data sources used as ground truth, when two or more sources have the same event, the dates

[2] `https://gadm.org`

[3] `https://archive.org/`

Table 5.1: Features extracted; p_i is the probability that tweet i is related to floods, as computed by an automated classifier.

Keys	
Name	Description
Date d	Day number
Region	Administrative region in GADM
Daily features (22 feat)	
Lang	0: English not an official lang, 1 or 2: English is 1st or 2nd lang
TOT	Tot number of Tweets on this day and this region
T00	In bucket Ta-b, number of postings having $a < p_i \leq b$
P00	In bucket Pa-b, fraction of postings having $a < p_i \leq b$
Lagged features (50 feat) computed over a moving window of 3 days	
T3P00	In bucket T3Pa-b, total number of postings having $a < p_i \leq b$ on days $\{d, d-1, d-2\}$
M3P00	In bucket M3Pa-b, fraction of postings having $a < p_i \leq b$ on days $\{d, d-1, d-2\}$; these add up to 1.0
A3P00	In bucket A3Pa-b, average fraction of postings having $a < p_i \leq b$ computed over the three days
D1T00	In bucket D1Ta-b, change in the number of postings having $a < p_i \leq b$ between day d and day $d-1$
I3T00	In bucket I3Ta-b, maximum increase in the number of postings having $a < p_i \leq b$ between day d and day $d-1$, or day $d-1$ and day $d-2$; this is always non-negative

are not necessarily the same. Some ambiguity is inevitable at testing time, but in the training data, we would like to learn only from unambiguous cases.

Our methodology associates the floods in our ground truth with specific time-spans (days) and regions. Then, we consider the dynamics of the discussion following the evolution of an event i, spanning between dates d_i^{start} and d_i^{end} in the ground truth. Figure 5.3 shows sample data for one such flood event. In Figure 5.3, the overall number of social media postings in the same region is represented by bars. To associate labels with days and regions in the training

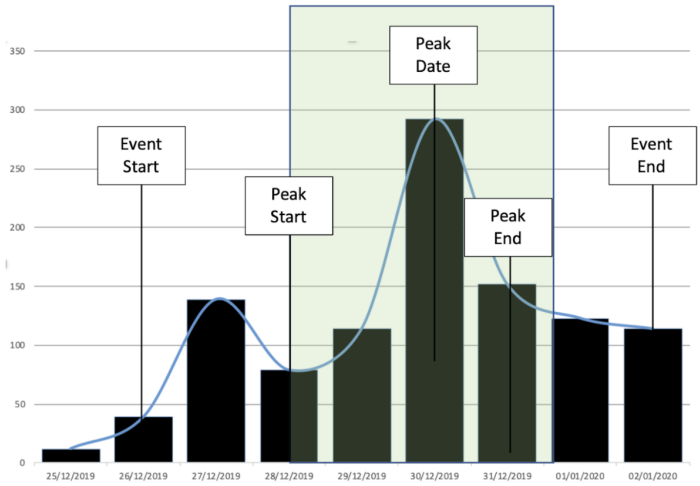


Figure 5.3: Overall volume of flood-related postings per days overlapping with a flood event i lasting eight days from 'Event Start' (d_i^{start}) to 'Event End' (d_i^{end}). The range that is labeled as *True* in the training data goes from d_i^b to d_i^e and reaches its peak at d_i^m (shaded area)

.

data, we consider a range of days within the series corresponding to the same region in which a flood was recorded in the ground truth. This range is created as follows:

1. We locate the day with the local maximum or peak $d_i^{(m)} \in [d_i^{\text{start}}, d_i^{\text{end}}]$ of social media activity within the days of the flood according to the ground truth.

2. We locate the beginning of that increase in activity, i.e $d_i^{(b)}$.

3. We set the end of the range to be the day after the maximum, $d_i^{(e)} = d_i^{(m)} + 1$, because in our observations there is almost invariably some conversation about the flood that remains in social media after the peak.

Then, we set all days within that region lying in the range $[d_i^{(b)}, d_i^{(e)}]$ to *True*.

A period of 10 days before and 10 days after the ground truth for the flood in the same administrative region is a grey area, due to the fact that increases social

media in activity may or may not be present. Hence, we set all days within $[d_i^{\text{start}} - 10, d_i^{\text{start}} - 1]$ and $[d_i^{\text{end}} + 1, d_i^{\text{end}} + 10]$ to *Undefined*. The remainder of the days in this region are set to *False*.

We also have to account for ambiguities in geocoding, which may associate floods in one administrative region with another administrative region in the same country. To preclude these from occurring in our training data, we remove labels from all other regions of a country where no floods are recorded in this period, in the period $[d_i^{\text{start}} - 5, d_i^{\text{end}} + 20]$. For regions where no floods are recorded within the entire observation period, we set all labels for all days to *Undefined*.

Finally, we considered only regions where English was an official language, and only days in which the total number of collected tweets was above 100. Our final training data contains 930 *True* or *False* rows corresponding to (day, region) pairs, of which 73 (or 7.9%) have the label *True*.

Building the model

We posed our research question as a binary classification problem, in other words detecting from the features extracted from postings whether these corresponded to a day and region with floods or without floods. We experimented with various classification schemes including Support Vector Machines, Multi-Layer Perceptron, and Random Forests. Random Forest (RF) classifiers yielded the best results.

We performed a grid search to optimize the learning parameters. For feature selection, we used an ANOVA F-Test, with the best performance obtained by selecting 40 out of the 73 features. The selected features using univariate feature selection with a classification function score covers most of the features classes in Table 5.1. They describe a combination of aggregated classes and average probability: 'T00-10', 'T10-20', 'T20-30', 'T30-40', 'T40-50', 'T60-70', 'T70-80', 'T80-90', 'T90-100', 'P00-10', 'M3P00-10', 'M3P90-100', 'A3P00-10', 'A3P60-70', 'A3P90-100', 'T3P00-10', 'T3P10-20', 'T3P20-30', 'T3P30-40', 'T3P40-50', 'T3P50-60', 'T3P60-70', 'T3P70-80', 'T3P80-90', 'T3P90-100', 'D1T10-20', 'D1T20-30', 'D1T30-40', 'D1T40-50', 'D1T70-80', 'D1T80-90', 'D1T90-100', 'I3T00-10', 'I3T10-20', 'I3T20-30', 'I3T30-40', 'I3T40-50', 'I3T70-80', 'I3T90-100', 'TOT'

For the RF parameters, we obtained the best results with 1,000 decision trees and a maximum depth of two levels for each tree, although we observed that

similar numbers of trees and depths did not yield a substantively different performance. We used three-fold cross-validation using two-thirds of the data for training and the remaining one-third for testing in each iteration.

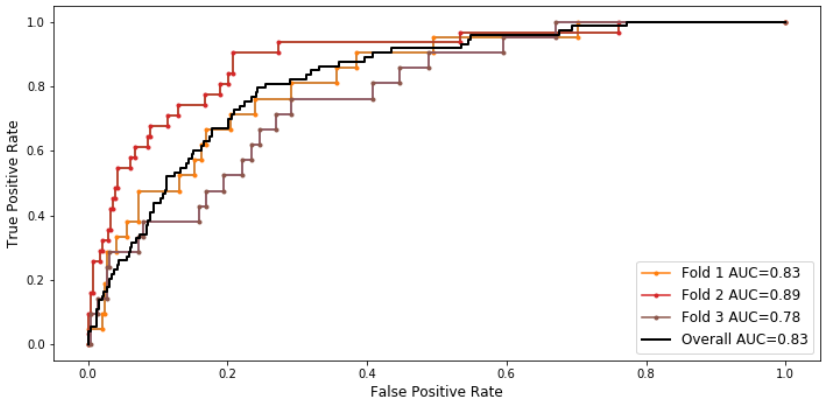


Figure 5.4: ROC curve of the obtained classifier.

Figure 5.4 shows the Receiver Operating Characteristic (ROC) curve representing the tradeoff between sensitivity (i.e. the true positive rate, or how good the model is at detecting real floods) and specificity (i.e. the true negative rate, or the model's ability of avoiding false alarms). An increase in sensitivity is accompanied by a decrease in specificity. We observed that the best model does not offer a high-sensitivity high-specificity classification. By setting a threshold for a positive (i.e. 'flood') classification of 0.2, which yield a good balance of precision and recall, we obtain an average combination of precision of 34% and recall of 41%. This implies that this model cannot be used independently of hydrological flood monitoring systems for detecting floods. However, as we will show in the following section, the GloFAS forecasting performance analyzed during the ground truth events would benefit from our classifier in terms of reducing uncertainty.

Leave-One-Out experiment

In the previous sections we calculated the accuracy of our model using statistic. The model presented, as described in Section 5.1, aims at an operational use together with global forecasting systems. In order to measure the accuracy of our

model in an hypothetical operational system, we need to define the hit-rate as the percentage of real events that our model could predict. We simulated such methodology using a Test/Train split supported by group definition. In our previous experiments we considered the rows labeled as '1' uncorrelated from each other. Although this is true, relying solely on random split could lead to bias test data towards a specific event (multiple rows). In other words, for this experiment we assigned an 'event_id' to each of the rows of the dataset and we grouped and isolated each time an event to define the Test dataset as this would be the case for future events. Hence, we consider how this model would perform in a 'leave-one-out' cross-validation scenario, and particularly, whether it can complement the forecasts of GloFAS.

Overall results

We first observed that the rows (days and regions) labeled *True* are correlated with each other, if they occur in the same region around the same time, as in the case that they represent the same flood. Hence, we cannot leave out one row, but instead must leave out an entire event.

Since we wish to perform a side-by-side comparison against an operational system for disaster risk reduction, we consider two possible outcomes: 'Hit' or 'Miss'. The former is when we trigger an alert for at least one of the days of a flood, while the latter is when we do not.

Table 5.2: Output of the leave-one-out classifier and GloFAS for 23 flood events.

Place	Country	Days	Result	GloFAS 20yr	Type of event
Suffolk	USA	5	miss	no river	Storm surge
Herkimer	USA	2	miss	10-20%	Heavy rain, flash floods
Maury Cty.	USA	2	miss	00-10%	Heavy rain, flash floods

The hit rate of the experiment for the 23 simulations done (i.e. one per each event in our training data) is 52%, which means that we capture about half of the flood events. Table 5.2 shows three cases from October and December 2019.

On a cautionary note, it is important to bear in mind that the main purpose of GloFAS is to forecast riverine floods, and therefore those events that are a combination of riverine, coastal and / or flash floods can only be compared to GloFAS forecasts to a limited extent.

Our methodology offer the obvious advantage to capture all types of flooding (coastal, flash flood, pluvial, etc.) with the same ML-trained model. We observed that in many of the cases our model indicates flood activity, although this must be considered in light of the computed average precision of 34% (described in the previous section), meaning that about one in three of the alarms generated by the model based on social media alone will correspond to a flood.

5.4. Case Studies

In order to better understand the hit-rate performances of the leave-one-out model, we analyzed some of the cases in detail.

Firstly, the floods of October 2019 in *Suffolk county* in the US were missed by our model:[4] 'Minor flooding was reported in parts of Suffolk County, New York. Roads were swamped and some buildings flooded'. In this case, the flooding can be considered minor, as neither fatalities nor injured persons were recorded in the ground-truth dataset. Our system would have missed this flood, and GloFAS indicated no signal.

Secondly, the floods of October 2019 near *Herkimer (Mohawk Valley)* in the US were also missed by our model:[5] 'According to a statement by New York Governor Andrew Cuomo's office on 01 November, over 240,000 homes were without electricity and nearly 60 roads were closed'. Although this event affected more people, the main impact was an electricity blackout. Our system would have also missed this flood, and GloFAS indicated a 10-20% chance of exceeding the 1 in 20-year discharge. At the peak of the storm, we found that 75% of the total postings were classified as not relevant to floods resulting in low values for features used by our model.

In both of these cases, the reason for which the GloFAS model had little or no signal was that the main driver of the flood was water from the storm and from storm surge (coastal flood), rather than water overflowing from a river.

Thirdly, the floods of December 2019 in *Gauteng and North West Provinces (near Hartbeespoort Dam)* in South Africa were captured by our model:[6] 'Hundreds or people have evacuated their homes. News 24 reported that one person died on 9 December when flash floods swept a vehicle from a low-lying bridge

[4] https://bit.ly/2UoOxg9

[5] http://floodlist.com/america/usa/halloween-storm-flood-october-2019-new-york

[6] http://floodlist.com/africa/south-africa-floods-gauteng-december-2019

close to Hartbeespoort Dam in North West Province, about 35 km west of Pretoria.'. In this case, GloFAS forecasted a probability of an extreme (1 in 20-year) event in the area, of 30-40%. Our test data features indicate more than 2,500 social media postings in one day, 40% of which were classified as highly relevant.

5.5. Conclusions

While the forecasting of floods using hydrometeorological models is possible within certain limits, many floods are not forecasted or are forecasted with only a low probability. Although comparing forecasts from a hydrological model would be fairer if only purely riverine floods were considered, only in one case of the flood events taken into consideration the computed probability of exceeding a '1 in 20-year' threshold forecast by GloFAS, was more than 50%. This is largely because forecasting systems are based on model simulations, meaning that they are affected by noisy signals due to many factors (e.g., noise in meteorological forecasts, missing data, and incomplete reference data).

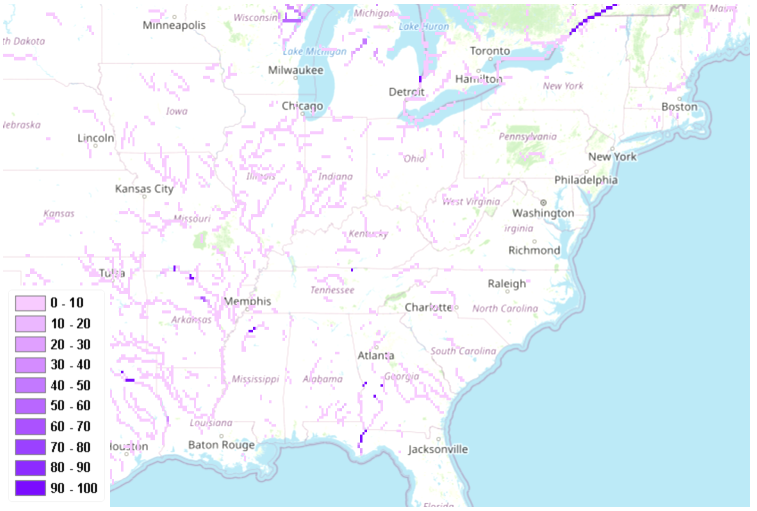


Figure 5.5: GloFAS map for April 2020, highlighting in purple portions of river basins that have a heightened probability of floods according to darkness: this happens in many areas in the US at the same time. (Better seen in color)

Both forecasters and emergency managers require tools that can help them to narrow down uncertainty. For example, Figure 5.5 shows the ensemble prob-

ability of exceeding the 1 in 20-year flood threshold in the US for April 2020, as forecast by GloFAS. As can be seen, GloFAS forecasts low to medium probability of a flood in many administrative areas.

The methodology proposed for leveraging our model in real-time is to keep the collection of tweets as described in section 5.3 running in the background and tasking the model classification on a daily basis. Since the model uses features with data from the previous two days, we think it can 'detect' a change in the conversation on social media as shown in Figure 5.3. The slope of the peak can be smoother or steeper according to the type of event, the classifier could be able to detect the event before its peak, which is associated with the highest impact. Our work provides an added value to the GloFAS hydrometeorological forecasts since it helps to reduce uncertainty and broadens the range of flood-types that can be detected

In particular, we are confident that the precision of our classifier in determining whether a flood is occurring in a specific area, could be improved by: (a) using our model as a trigger for a more focused real-time data collection, where city names are used instead of flood-related keywords, and (b) setting threshold levels for the ratio between tweets classified as 'most likely flood-related' and those classified as 'likely flood-related', for a real-time aggregation of data. In the latter case, the ratio can act as an indicator to filter out noise caused by trending topics in the specific area (i.e., sport events, celebrities, politics, etc.), where we expect to have more tweets unlikely to be about floods.

Another potential improvement to be addressed in future work, concerns the normalization of features. In our study, we normalized the features of the training dataset using the number of expected data at the national level. However the adoption of a particular social media platform may vary between regions within the same country. Another practical issue that merits further investigation is how to handle multiple crisis events (not necessarily all natural hazards) happening in the same country. When there is a strong trending topic, we observed that floods might receive less attention from mass media and from those members of the public who are not directly affected, which reduces the strength of the social media signal.

Our research has demonstrated that the methodology proposed can complement a global flood forecasting system. One aspect which we have not yet addressed is the potential link and research related to the concept of lead time of the forecast. In other words, so far we have analyzed the additional value of

the model on the day of the event, while there is still potential improvement in considering the forecast of an event, specifically when it builds over several days.

Reproducibility
All of our code, as well as data to reproduce the results on this chapter, are available for research.[7]

In the next chapters we are going to explore the potential of social media messages as a source for detailed information that could result in quantitative information or in actionable information. This step introduces additional challenges such as trustfulness and validation of data that we address in Chapter 6 and Chapter 7.

[7] https://zenodo.org/record/4274495

Part III

Social Media Analysis for Impact Assessment

Chapter 6

SOCIAL MEDIA AS A WAY TO COMPUTE FLOOD EXTENT MAP

In this chapter we proceed in the pursuit of an answer for *RQ4*, specifically we experiment if it is possible to leverage real-time information from social media, fusing it with digital surface models derived from EO data to provide a fast estimate of the flood extent.

6.1. Introduction

Citizens constantly use social media, including for broadcasting content during disasters and emergencies. The vast amount of such data originating from the public can be used to provide access to timely and relevant information, offering additional decision-making support to emergency managers. Images extracted from social media can be vital in the immediate aftermath of an event when authoritative data and products based on Earth Observation (EO) are not yet available. Mapping ground truth information is crucial for the early assessment of impacts, in terms of their intensity and spatial distribution. As part of the On-Demand Mapping component of the CEMS, satellite imagery and other geospatial data are used to provide mapping service free of charge for natural disasters, human-made emergencies and humanitarian crises throughout the world. Only authorized users such as civil protection, entitled emergency response organizations or international charters can activate the service. The maps are avail-

able in two temporal modes: RM, and Copernicus Risk and Recovery Mapping (RRM). The former provides geospatial information within hours or days of the activation following a disaster, while the latter provides geospatial information supporting disaster management activities not related to immediate response.[1]

Floods represent 36% of activations of the CEMS RM service.[2] Flood extent is normally delineated using both Synthetic Aperture Radars (SAR) satellite images (especially those from the Sentinel-1 satellite) – useful for detecting water-covered areas even at night or in the presence of clouds - and image data from optical EO sensors, which allow the identification of damages for impact assessment. Unfortunately, the effectiveness of such satellite image analysis is of limited use in urban areas, at the point that these are commonly masked and not analyzed. For optical sensors, the limitations are due to shadows cast by buildings, trees or narrow streets. For SAR sensors, the side-looking viewing geometry and the multiple scattering in built-up areas do not allow the presence of water to be properly distinguished.

On 13th November 2019, the mayor of the Italian city of Venice declared a state of emergency after an exceptionally high tide, recorded as the worst in 50 years, flooded the city.[3] A deep cyclonic circulation had affected the Mediterranean area the previous day, resulting in severe weather over the Italian peninsula. One of the most affected areas was North-Eastern Italy, particularly the Friuli Venezia Giulia and Veneto regions. A full moon (+26 cm surge), combined with the exceptionally high level of the Mediterranean sea in November 2019, and a deep small-scale atmospheric pressure moving rapidly northward and passing over the Venice lagoon just west of the city (+30/35 cm surge), led to a high tide with a maximum recorded value of 189 cm on 12th November at 10:50 p.m. (hereafter we refer to local time) - the highest recorded since a similar event in 1966 [11]. According to the altimetry available to the municipality of Venice, as a result of an *'Acqua Alta'* (high tide) of 189 cm, about 82% of the public pedestrian traffic areas were flooded. The impact on the city was dramatic, with two fatalities in the Pallestrina neighbourhood, severe damage to the crypt of the San Marco basilica, three ferries sunk[4] and 2,494 claims for economic loss

[1] https://emergency.copernicus.eu/mapping/ems/service-overview

[2] https://emergency.copernicus.eu/mapping/ems/rapid-mapping-portfolio

[3] https://www.nytimes.com/2019/11/13/world/europe/venice-flood.html

[4] https://www.ilgazzettino.it/nordest/venezia/acqua_alta_

of totalling 9 million euros in damages by residents and businesses, in the initial weeks following the event.[5]

The RM service of the CEMS was activated by the Italian National Civil Protection Department on 14th November at 01:15 p.m.[6] as other exceptional high tides were forecasted for the following day. Among the set of available images from space satellites, the closest to the time and area of the event was acquired by the GeoEye optical satellite sensor on 14th November at 11:13 a.m. Flooded areas were not adequately visible due to the narrow streets of the city and the high off-nadir acquisition angle of the image (not vertical). Thus the optical imagery was complemented by ancillary data: tide-levels combined with contour lines. The flood extent delineation that was delivered on 15th November at 09:18 p.m. shows the detected situation when the tide recorded had fallen to about 115 cm, with an estimated flooded area below 30%. This significantly failed to capture the maximum extent flood later reported by the municipality.[7] What we learn from this is that, the use of remote sensing to delineate flood extent in a city can be incomplete or inaccurate, especially for fast-developing events such as urban floods.

Between 12th and 14th November 2019, we collected posts on Twitter using specific filters related to the Venice flood event. In order to fill the gap in information between the immediate aftermath of the event and the moment when authoritative data were available, we delineated a potential flood extent based on images that were classified as relevant to the event and data available for free to the public. In this way we were able to estimate a maximum flood extent similar to that recorded and validated by authorities. This work presents a scalable methodology for combining deep learning models for image classification with global or local Digital Elevation Model (DEM)s and other geospatial information, for a near real-time delineation of the flooded area. The results show how the use of social media information for urban floods can complement EO data and can help to improve situational awareness. We present a fast methodology complementing both approaches. It offers an additional source of in-situ data that can serve as input for hydraulic models and provides a reference layer for filling spatial and temporal gaps in EO-based products not available during urban

allarme_notte_venezia-4858433.html

[5] hhttps://live.comune.venezia.it/it/dati-richieste-danni-acqua-alta-12-novembre-2019-venezia

[6] https://bit.ly/3AzHJA4

[7] https://bit.ly/3dMpzkX

floods.

The experiments presented demonstrate how such a layer could estimate a flood extent map within the first 24 h of an urban flood. In addition to the benefits mentioned above, the proposed approach uses data that is free of charge except for the geocoding step, therefore having a low economic impact compared to the cost of EO imagery acquisition and analysis.

In the the remainder of this chapter, related research work is presented in Section 6.2, technical details of the data collection, the methodology and the experimental results are presented in Section 6.3, and finally, a general discussion of the potential of social media analysis for assessing floods extents and future developments is presented in Section 6.4.

6.2. Related Work

Recently, Pastor-Escuredo *et al.* (2018) suggested integrating social media data into a framework consisting of authoritative and non-authoritative data for assessing impacts of a disaster. In their work, the function of social sensors regarding mapping the extent is limited to mobile phone use for mobility detection, which is a valuable source of data but makes the methodology hard to scale for cases where such information is not accessible.

Last year Xiaoyan *et al.* (2021 introduced an innovative approach to estimating flood-affected populations, providing high-resolution impact information. The described case study shows that considering mobility patterns during assessment can improve the precision of disaster estimation. Inundation locations and roads blockage are detected by combining flood hazard maps with social media data, thus applying historical statistical data and real-time user-generated data. Social media data regarding inundations were obtained from the official Weibo account of the Wuhan Traffic Management Bureau.

Our methodology can be applied in near real-time, overcoming the timeliness limitation of post-event analysis. Our work also provides a methodology that can be reproduced in different cities, as it uses open and globally available data.

Several research studies have shown the potential, opportunity and limitations of satellite images and radars for natural disaster analysis. Over the years, increasingly powerful methods and sensors have been launched on satellites, to avoid weather-related signal attenuation. Commonly used technologies include LIght Detection And Ranging (LIDAR), which can detect the altitude of objects from a long distance, and SAR, which creates two-dimensional images using a

signal frequency unaffected by light conditions (day or night) or cloudy weather. Mason *et al.* (2018) studied a method to detect floodwater in urban areas with a SAR simulator in conjunction with LIDAR data. The method allows predicting areas of radar shadow and layover in the image caused by buildings and taller vegetation. The results indicate that flooding can be detected in an urban area with reasonable accuracy. However, the algorithm design assumes that high-resolution LIDAR data are available for the area under analysis. In 2021 the same authors use open-access Sentinel-1 SAR data, the World-DEM digital surface model (DSM), and open-access World Settlement Footprint data to identify estimates of flood levels in urban areas locally. Their method searches for increased SAR back-scatter in the post-flood image due to double scattering between water (rather than non flooded ground) and adjacent buildings and reduced SAR back-scatter in areas away from high slopes. The method reports high accuracy in moderately dense housing areas, while the accuracy decreased in dense housing areas when street widths are comparable to the DSM resolution. Lin *et al.* (2019) employ SAR intensity time-series statistics to create a flood probability map. The resulting extent is selected by applying a global cutoff probability of 0.5. However, smooth surfaces like asphalt roads, SAR shadow, aquatic plants, and soil moisture changes introduce inaccuracies in the prediction. Furthermore, the long time for processing images to build the SAR intensity time-series statistics makes the methodology unsuitable for real-time deployment. The methodology proposed in this research uses social media information fused with digital surface models as sensors for detecting ground truth in near real-time to reduce uncertainty and contribute to solving the issues related to EO technologies.

The so-called 'digital divide', and a lack of resources especially in vulnerable area more impacted by global warming [101] [19], have focused the attention of scientists and crisis responders on research tools and methodologies for flood risk management at a global scale. Flood risk assessments for cities produced using Global Digital Elevation Models (GDEMs) are likely to over-predict risks. Past studies found variability in the accuracy of models using different GDEMs, and all substantially estimated higher impacts than the DEM produced from aerial LIDAR [61]. As the world's cities grow, the importance of accurately understanding flood risk has become a high priority. GDEMs enable flood risk assessments to be undertaken globally, defining standard methodologies allowing data integration. Uncertainties in flood risk assessment using GDEMs need to

be addressed and reduced in near real-time by local and national authorities and communities, to prevent misinformed decision-making.
International organizations have put in place emergency management services with the aim of providing support to crisis responders who are in need of resources. The European Union's Earth observation programme, called Copernicus, offers information services that draw from both satellite EO and in-situ (non-space) data. As part of Copernicus, the CEMS supports local authorities and communities needing information to develop environmental legislation and policies or to take critical decisions in the event of an emergency, such as a natural disaster or humanitarian crisis. The Early Warning systems and On-Demand mapping components of the CEMS produce flood hazard maps that have been developed using hydrological and hydrodynamic models, driven by the climatological data of the European and Global Flood Awareness Systems (EFAS [93] and GloFAS [4]). All maps are in raster format with a grid resolution of 100 m (European-scale maps) and 30 arcseconds (global-scale maps). These maps can be used to assess the exposure of population and economic assets to river floods, and to perform flood risk assessments.
Our research is aligned with and complementary to the previous research on providing near real-time information during floods, particularly in densely inhabited areas. A key advantage of our contribution lies in the advantage of the real-time aspect of social media data, together with physical model and EO data. Our research can answer the following important topical question: *Is it possible to leverage real-time information from social media, fusing it with digital surface models derived from earth observation data to provide a fast estimate of a flood extent?*
The answer to this question partially responds also to *RQ4* presented in Chapter 1. The proposed workflow uses social media information to find flooded points (latitude and longitude). It then infers the spatial extent of the flooded area operating a vertical data interpolation based on digital surface grid-based information. Assuming that the same amount of rainfall fell on the city, if point A is flooded and point A is in a higher grid-cell than point B, and their grid-cells are close, we assume that both grid-cells are flooded. The tool's accuracy is determined based on the grid-cells estimated as flooded against the flooded grid-cells provided as reference.

6.3. Methods

In order to analyze the quality of flood extent mapping based on social media information during urban flooding, we have carried out two experiments, both in the context of the flood that hit Venice in 2019, as was described earlier in Section 6.1. In this section, we describe in detail the data collection, the applied methodology, and the results that were obtained from both experiments.

Weather forecasts
On 12th November 2019, a deep cyclonic circulation affected the Mediterranean area, which resulted in severe weather over the Italian peninsula. In Venice, the extreme weather condition produced a high tide whose maximum recorded value was 189 cm at 10:50 p.m.. At the start of the event, on the morning of 12th November, the municipality emergency managers expected a high tide peak of 170 cm at 11:00 p.m. and another peak of 160 cm for the following morning at 10:30 a.m. Several warnings were issued, schools were closed and travel restricted, and plans for emergency response were activated.

Social media
One of the key elements of the methodology proposed here is the consideration of social media as valid ground truth information. We searched for messages posted on Twitter from 12th November at 01:00 a.m. to 14th November at 01:00 a.m., either geocoded inside Venice island or mentioning the keywords 'Venice' or 'AcquaAlta'. We collected roughly 75,000 tweets, 14,000 of which contained pictures.

Digital models
Collected Social Media data are then combined with digital models representing the surface of the city. In order to study the quality and the scalability of the methodology proposed we used several DEMs:

1. SRTM (Shuttle Radar Topography Mission) is a global research endeavor that yielded nearly- global DEM with a 30 m resolution. SRTM data covers the globe and is free of charge (though registration is required).

2. The Copernicus DEM EEA-10 instance (hereafter referred to as Copernicus DEM), available free of charge from the European Environmental

Agency[8], is a Digital Surface Model (DSM) which represents the surface of the Earth including buildings, infrastructure and vegetation with a resolution of 10m. Since the Copernicus DEM includes building, to reduce as much as possible the error in the elevation of the control points, we adjusted the values of height extracted. Let $PBUILT(lat, long)$ represent the 0 to 1 share of buildings in a 10 m cell at $(lat, long)$ according to the GHSL built-up-area product[9].
$ELEV(lat, long)$ is the original elevation value from the Copernicus DEM model Then the adjustment formula that provide the corrected elevation value $ELEV'(lat, long)$ can be described as:

$$ELEV'(lat, long) = ELEV(lat, long) - PBUILT(lat, long) \cdot ELEV(lat, long) \tag{6.1}$$

While the formula works as it is for cities at sea level, to reproduce such adjustment in higher urban areas, we should lower the elevation $ELEV(lat, long)$ values by subtracting the average altitude of the site or the height of the nearest cell with a near-zero share of buildings before applying the formula.

3. TINITALY [95] is a seamless DEM of the whole Italian territory. This DEM, which was produced starting from separate DEMs of single administrative regions of Italy, is freely available with a 10 m grid-cells. TINITALY is published with a CC BY 4.0 license and can be used freely, even partly, but it must be cited.

Methodology
The production of a near real-time flood extent map is carried out in four separate steps: collection of tweets; extraction of social media flood points; interpolation of social media flood points; production of flood extent. Figure 6.1 depicts the four steps while their details are provided in the remainder of the section.

Firstly, we collected tweets as described above. For conducting the experiments we relied on the so-called historical search available only to Twitter PowerTrack API users. Nonetheless the collection of tweets in real-time for replicating the experiment does not require any special data access, since tweets can be

[8] `https://www.eea.europa.eu/`
[9] `https://ghsl.jrc.ec.europa.eu/ghs_bu2019.php`

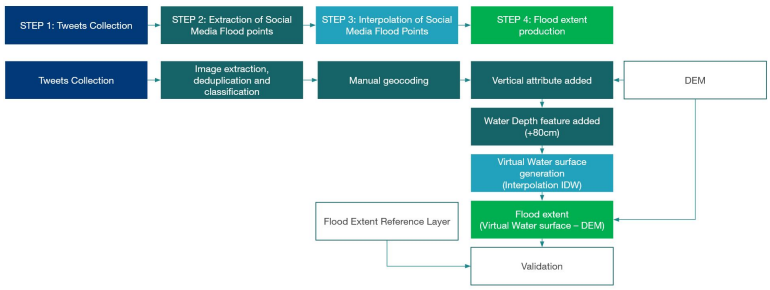


Figure 6.1: (best seen in color) The flood extent production consists of 4 steps: Indigo - Tweets Collection; Teal - Extraction of social media Flood points; Blue - Interpolation of social media Flood Points; Green - Flood extent production.

filtered from the publicly available Twitter streamer.[10]

The work aims to map the flood extent in a city, and therefore it is of the utmost importance that we geocode information as precisely as possible. Although Twitter enables users to post tweets with their current locations (longitude and latitude), only an average rate of 0.85%–3% tweets are being geocoded per day [109] in the USA, where Twitter is most used. Thus, we need to complement the geocoded dataset using other techniques to ensure scalable global products. Past works proved that location mentions were useful to geocode information [97]. However, after a first test using several hundred tweets, we understood that we could not rely solely on automated NLP for a precise geocoding. Figure 6.2 shows one of the many examples of how a location mentioned in a text does not correctly represent the location of the information. Previous works [2] found that tweet images contained more damage-related information than their corresponding text. Thus, we opted to extract social media images rather than text as they could be better inspected for geocoding within the city and for the extraction of crisis information about the event. During the two days we collected 14,000 images, resulting in 10,000 images after duplicates were removed using a tool for checking and deleting near-duplicate images based on perceptual hash.[11]

The second step of the methodology is the identification of social media flood

[10] https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time

[11] https://github.com/knjcode/imgdupes

Figure 6.2: Example of the wrong facility identified by NLP, as it was mentioned in the tweet text 'flood reached the maximum peak cm at Punta della Salute'. The correct location is derived from the tweet picture

points to be considered as a control point for the flood mapping activity. Once the tweets were collected, we used a Convolutional Neural Network (CNN) model for disaster image classification [86] to classify the images of a flooded location. The model assigns a probability of an image to belong to one of five classes (Flood, Wildfire, Storm, Earthquake, Other). In particular we set a threshold of flood probability equal to 0.9 for identifying those relevant to the event. We found 2,302 images depicting flooded areas, some examples of which are shown in Figure 6.3. The vast majority of the relevant images were then geocoded manually. The location of the images was based on the identification of recognizable Points of Interest (POIs) like shops, monuments, street names, bridges, public transport stops, and their comparison with Google Street View. When the image showed a wide area, such as a square photographed from a building, the location of the image was placed in the flooded area, i.e. the center of the square. When, in a flooded street, a shop could be identified, the location was placed in front of it. The annotation was verified by a contractor of CEMS, who found that 97% of images were properly geocoded.[12] Since this work proposes a scalable

[12]the report, written by Trabajos Catastrales S.A., is available upon request.

Figure 6.3: Examples of images classified as relevant and manually geocoded.

methodology for detecting a potential flood extent map, manual geocoding has been performed considering the time component. As already described earlier in Section 6.3, only an average rate of 0.85% to 3% of the tweets are originally geocoded. To speed up the geocoding of images, we used an NLP tool[13] to extract mentions of place names from text and leveraged such information to locate the flooded point. However, we have not monitored the time consumed for the single geocoding for the experiments. Thus we cannot estimate how much such automated pre-processing contributed to speeding up the geocoding. Images that could not be geocoded or that were clearly referring to weather conditions in areas outside Venice, were excluded. Finally, after the manual geocoding we could identify almost 800 social media flood points, and 265 unique points. The focus of our research is to support the development of a new product that could be made available within the first 24 h after an urban flood happened when neither EO-based nor authoritative maps are available due to technical challenges. The feasibility of such a product is confirmed if we consider that the manual geocoding process, done by non-local personnel using Google Street View, took 6 h (one person). If needed, we can assume that a service provider could allocate more resources to this task during an actual case. Furthermore, manual geocoding can be quickly done during a real crisis using crowdsourcing to leverage the help of local digital volunteers, coordinated by practitioners and emergency-oriented volunteers, such as Virtual Operations Support Teams. (VOSTs)[14]

Figure 6.4 shows a comparison between tweets originally geocoded on a no-flood day and the tweets identified as social media flood points. We divided the city of Venice into a grid of 50x50 m cells and counted the tweets in each cell.

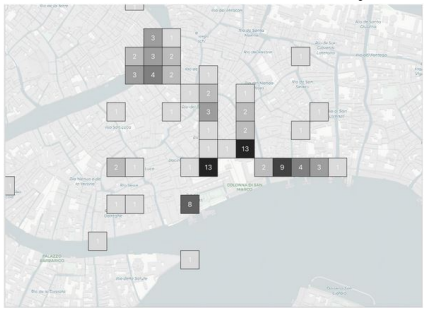
[13] https://github.com/deepmipt/DeepPavlov

[14] http://vosteurope.org

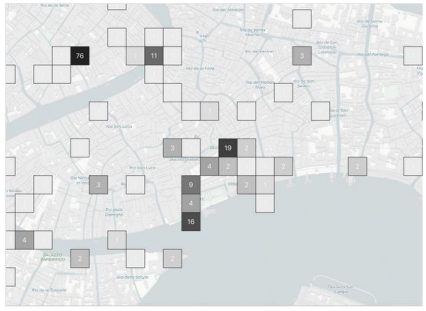
We carried out the comparison in order to ensure that the experiment could be scaled to other less-visited cities. The random scattered distribution of geocoded tweets on a no-flood day clearly demonstrates that the image classification step leads to an unbiased distribution. In effect we demonstrated that more pictures taken does not correlate with more flood points.

Filtered tweets on a flood day

Unfiltered tweets on a no-flood day



Number of filtered tweets on a flood day per 50x50 m cell



Number of unfiltered tweets on a no-flood day per 50x50 m cell

Figure 6.4: Geographical distributions of filtered and unfiltered tweets. (Note that the darker the cell, the higher the number of tweets in the cell)

The social media flood points, defined by latitude and longitude coordinates, were given a vertical attribute by sampling the available DEM. We created three datasets of points, one per DEM as defined in at the beginning of the chapter. A default water depth (DWD) of 80 cm was added to the ground vertical component based on visual inspection of the images (i.e. water above knees of passengers, doors). We considered such approximation done in other works [8] acceptable to contain the time of processing. A better estimate of the water levels in each

image would undoubtedly lead to more accurate results, but it would not apply to a semi operational process. This aspect is further discussed in Section 6.4. In riverine flood hydrology, flood simulation models are used in combination with terrain analysis to detect the flow of water towards lower or unprotected areas. Control points can be used to detect water-levels in statistically generated flood hazard maps. The main driver in the case of urban floods is the amount of rain falling on an area combined with malfunctioning man-made artifacts (i.e. sewage, buildings, roads). Thus, instead of focusing on the water flow, the third step of our methodology is to create a virtual water surface interpolating the social media flood points prepared in the second step to estimate values at other unknown points. We favored an Inverse Distance Weighted (IDW) method, where the sample points are weighted during interpolation such that the influence of one point relative to another declines with distance.

The fourth and last step of our methodology is the identification of the flooded area. This is obtained comparing for each point of our map the DEM and the virtual water surface generated at the previous step. When the water surface is higher we assume the cell is flooded.

Experiments
During the interpolation, we tested different values of the coefficient IDW-P, to create a few different surfaces and adjust this parameter to suit our analysis. A larger coefficient means it takes a larger distance for the values of the surface to become dissimilar from nearby points. A small coefficient means the values of the surface will quickly change as distance increases.

Experiment 1: we interpolated the social media flood points with several weighting values to determine the best accuracy for the maximum extent of the flood. We created an elevation reference layer (altimetry) using the contour lines relative to the elevation of the pavement of the historic centre of Venice with respect to the median sea level, whose accuracy is 1 cm vertical and 2 cm horizontal.[15] According to authoritative sources (as mentioned in Section 6.3), a maximum level of 189 cm was recorded at 10:50 p.m. on 12th November 2019. Thus, by selecting only the points of the elevation reference layer below that value, we were able to define the maximum flood extent to use as reference map. All the

[15] `http://smu.insula.it/index.php@option=com\_content\&view=article\&id=15\&Itemid=111.html`

layers were transformed to grid-based maps (rasters) during the experiments, and statistics were done on such grid cells. In order to compute Precision and Recall for the experiment, we considered four types of result (True Positives, False Positives, True Negatives, and False Negatives), as described below:

1. **True Positive** values of the cells that are detected under water from both our methodology and authoritative data (altimetry).

2. **False Positive** values of the cells that are detected under water from our methodology but not according to authoritative data (altimetry).

3. **True Negative** values of the cells that are not detected under water from both our methodology and authoritative data (altimetry).

4. **False Negative** values of the cells that are not detected under water from our methodology but they are, according to authoritative data (altimetry).

Table 6.1 displays the results of the methodology proposed using several values of the weighting coefficient IDW-P for the interpolation of the social media flood points. We ran the majority of simulations with the Copernicus DEM, because we are convinced that the CEMS could benefit from this work. All the simulations were compared against the maximum extent identified by reference contour lines, as we used the social media flood points collected over the whole period of the event. The first row of Table 1 outlines a trivial experiment where we assumed the entire city was flooded. Given the magnitude of the event, it seems that such an assumption brings good results. Thus, for clarity, Table 6.1 reports also the Matthews Correlation Coefficient (MCC), that ranges in the interval [-1,+1], with extreme values -1 and +1 reached in case of perfect misclassification and perfect classification, respectively, while MCC=0 is the expected value for the coin-tossing classifier. According to [12], this coefficient shows more reliable evaluations versus overall accuracy (OA) and the F1 score, particularly on imbalanced data-sets, such as in our case where the majority of the pixels were flooded. The values in bold represent the best score for each column, maximized in case of true values and minimized in case of false values. We notice that among the runs with the Copernicus DEM, the IDW-P coefficient carrying the best results is 10. Specifically, it detects the highest number of true negative cells, which is valuable information given that almost all the city was flooded. The last two rows of Table 6.1 represent the methodology's simulation using the

best IDW-P (10) but with the other DEMs. The local DEM, TINITALY, gives by far the best results, especially in detecting the true positive values. An interesting feature to emerge was how the SRTM DEM offers an excellent alternative to the Copernicus DEM, although with a lower spatial resolution (30 m against 10 m).

DEM	IDW-P	TN (%)	FN (%)	FP (%)	TP (%)	OA (%)	MCC
AllFlooded	No	0	0	8.91	91.09	91.09	0
COP	2	4.45	38.03	4.46	53.06	57.51	.047
COP	4	5.65	36.85	3.26	54.24	59.90	.132
COP	10	**5.97**	36.58	**2.94**	54.51	60.48	.154
COP	15	5.82	36.14	3.10	54.94	60.76	.147
COP	20	5.80	35.80	3.11	55.28	61.09	.149
COP	25	5.78	35.57	3.13	55.52	61.30	.149
COP	30	5.78	35.53	3.14	55.55	61.33	.149
SRTM	10	5.78	31.04	3.13	60.04	65.82	.182
TINItaly	10	3.37	**7.56**	5.54	**83.53**	**86.89**	**.269**

Table 6.1: Accuracy comparison between interpolations made with different weighting coefficient IDW-P. (Note: DEM = digital elevation model; IDW-P=weighting parameter; TN = True Negative; FN = False Negative; FP = False Positive; TP = True Positive; OA = Overall Accuracy; MCC = Matthews Correlation Coefficient

The thematic validation was performed by calculating pixel-based confusion matrices from which we can extract the overall accuracy (OA) for the different IDW-P values and DEMs. Figure 6.5 shows an overview of the validation for IDW-P=10 coupled with the Copernicus DEM.

Areas highlighted in green represent the pixels where there is agreement between the estimated flood and the reference layer (TP or TN). In purple we represented the omission errors (FN) and in orange the commission errors (FP). The dots in yellow represent the control points. It can be seen how the main omissions are in the areas where there are no control points. This can be explained as in these areas, due to the weighting parameter, the interpolation layer does not report the presence of water. Indeed its value, despite being higher than zero, does not reach optimal values like the other coefficients (that range between 0 and 1), demonstrating the difficulty of our method to detect the non-flooded ar-

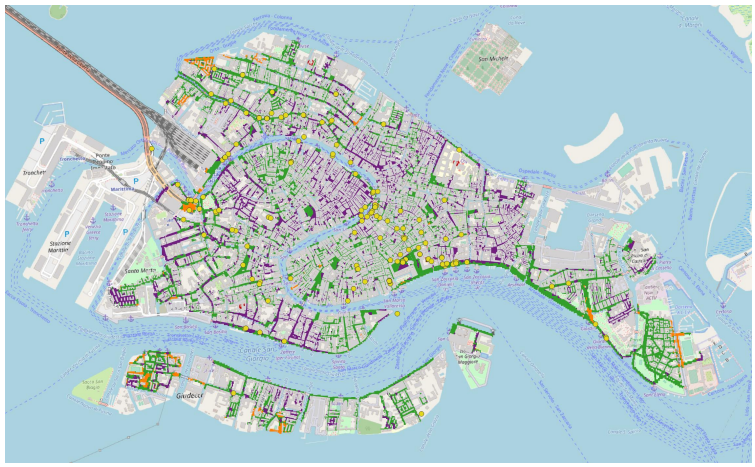


Figure 6.5: Overview of the validation for IDW-P=10 and Copernicus DEM (best seen in color).Areas in green show agreement between estimated flood and reference layer (TP or TN). Areas in purple show the omissions (FN) and in orange commissions (FP). Dots in yellow represent the control points.

DEM Copernicus 10m DEM SRTM 30m DEM TINITALY 10m

Figure 6.6: Overview of the for IDW-P=10 and Copernicus, SRTM, and TINITALY DEMs

eas (high value of FN) simply because we use control points only where images show an inundation.

Experiment 2: we performed a second experiment simulating a real-time scenario where only forecasts but no authoritative data about the high-tide were available. Assuming our methodology with an IDW-P value of 10 gives the best approximation, we produced a flooded surface map using only social me-

dia available the first day until 13th November at 01:00 a.m., and compared it with the reference altimetry below 140 cm forecasted by meteorologists for 12th November. We computed the true and false values comparing simulations with two different contour lines. The first with flood expected for altimetry data below 140 cm and the second with the reported (after the event) value of 189 cm. Table 6.2 shows the best results in terms of true values, overall accuracy and MCC.

DEM	CL	IDW-P	TN (%)	FN (%)	FP (%)	TP (%)	TN+TP	OA (%)	MCC
COP	189	10	5.42	17.16	3.49	73.92	**79.35**	**.793**	**.286**
COP	140	10	11.90	10.69	23.54	53.87	65.77	.658	.194

Table 6.2: Accuracy comparison between interpolations made with best IDW-P. with forecasts and 24 h social media for the day November 12 2019. DEM = DEM; IDW-P=weighting parameter; TN = True Neg; FN = False Neg; FP = False Pos; TP = True Pos; OA = Overall Accuracy; MCC = Matthews Correlation Coefficient

We can see from Table 6.2, based on the values obtained from the social media flood points after 24 h, we could already assume that the forecasted values of 140 cm were exceeded by far. The flood extent map produced after 24 h was closer to the one created by the municipality with a reference water level of 189 cm as reported afterwards.

6.4. Conclusions

The accuracy we obtain when determining whether a cell is flooded or not, using maps that are freely available for the entire world, is 61.3% at the 10 m resolution and 65.8% at the 30 m resolution. Using a more detailed, country-specific map, we arrive to 86.9% resolution (Table 6.1). We can say that our experiment answers part of our research question ('Is it possible to leverage real-time information from social media, fusing it with digital surface models derived from earth observation data to provide a fast estimate of a flood extent?') in the affirmative, using both local DEM data and the freely available Copernicus DEM. The experiments demonstrate that it is possible to estimate quickly urban flooding extent using freely available resources at a fraction of the cost usually needed for satellite image processing. Such a methodology can help to resolve the issues presented in Section 6.2 of this chapter, particularly regarding

the problem of EO-based products in an urban context and the difficulty of capturing the development of the flood events. The target accuracy for a mapping service like Copernicus Rapid Mapping is higher than 80%.[16] Accuracy in Urban areas appears to be usually lower, therefore most of the times urban floods are not analyzed. Furthermore, during a recent CEMS workshop[17], practitioners expressed interest in having rapid exposure assessment while waiting for the first RM product. It appears that while with a high-resolution DEM our tests show a high level of accuracy, even the worst-case accuracy achieved during our experiments could be valuable information for situational awareness in the event's immediate aftermath. For instance, local emergency responders might use the flood extent map as a starting product and refine the map by gathering complementary in-situ data based on their expertise and knowledge of the distribution of the city's critical infrastructures.
While the accuracy may vary depending on the resolution of the DEM, we are aware of the challenges presented by the geographical scalability of the methodology. Ranging in (i) the availability of tweets, (ii) the availability of local expert volunteers, (iii) the availability of technological tools for geocoding of the social media control points, affect the feasibility of the mapping product. Our future work will aim at defining classes of cases and experimenting the timeliness and applicability of the methodology. The classes should range between the two ends:

- Flood extent map feasible: high number of tweets, high presence of local volunteers, google street view, or other similar tools available (i.e. mapillary[18]), and a high-resolution DEM. The time needed for the products is less than 24h and the accuracy is high.

- Flood extent map impossible: few messages, low presence of local volunteers, no digital images to support geocoding.

An operational service could then estimate the applicability of the methodology within the first 24h and decide whether to add it or not to the data available

[16] https://etendering.ted.europa.eu/document/document-old-versions.html?docId=44850

[17] https://emergency.copernicus.eu/mapping/ems/cems-week-2021-conclusions-community-insights-and-service-evolutions

[18] https://www.mapillary.com/

to the crisis responders. If we consider the case of the CEMS activations for urban floods in European cities, with the support of EU-wide local experts (either volunteers such as VOST EU or contractors that provide the services for a fee), where google Streetview is almost overall available, we suggest the systematic use of the methodology as a product to complement disaster risk management services such as CEMS. At the same time, the tool needs further analysis for scalability and future applicability to more cases.

The accuracy of CEMS maps is routinely verified and validated; a contractor of the CEMS repeated this experiment with a more accurate elevation layer and used just a subset of 77 (out of 265) images. They achieved 76% accuracy, higher than our experiments, suggesting that as long as the elevation layer is accurate the number of required labeled images does not need to be very large.[19] Therefore, we can safely assume that it is possible to provide an estimation of the flood extent map since Twitter streamer filtering can collect tens of tweets within the first hours after an event. A possible solution for scalability could be that a set of maps is automatically produced every 6 h with data available. The geocoding could be contracted to a service provider to allocate resources case by case. CEMS Service Providers could inspect such products before being released, as currently done for other EO-based products.

Although much social media information is textual, it is challenging to geocode precisely (within a few meters) information, using the locations mentioned in the text. These often refer generically to a road or a place such as a square or a large facility. The use of relevant images from social media is crucial as it offers a better possibility of placing social media flood points. The methodology proposed relies on the automated classification of images to facilitate the identification of informational data. One limitation which we are already planning to address, is the possible combined effects of rain-driven and riverine floods in cities with a mix of built-up areas and river catchments. In this specific context, the methodology described could integrate an additional parameter for the interpolation of control points, namely the Height-Above-Nearest-Drainage (HAND) [67] terrain model, that takes into account groundwater dynamics.

The experiments described here for the case of Venice show very high precision, because almost all of the city was flooded. For this reason we also investigated other measures, and we aimed to derive results as real positive and real negative numbers. We also analyzed results using the mapping validation tech-

[19]This report, written by Trabajos Catastrales S.A., is available upon request.

nique. Our methodology searches for and utilizes social media flood points, thus maximising agreement in terms of true positive values, rather than minimising omission errors, as specified in Section 6.3. Future work could also focus on optimizing the search for non-flooded areas through the inspection of images.

Finally, it is worth mentioning that the social media flood points presented in this chapter can also be evaluated as a potential input layer for hydraulic models to reduce uncertainty introduced by weather forecasts. All the data and code used in this chapter will be available in a public repository with camera-ready article describing it (at the moment of writing it is accepted but the conference will be later in 2022).

Further research should focus on improving the automation of data filtering and reducing the overall time and resources used for geocoding data. One solution would be combining textual and visual information to support the manual geocoding of information. The following chapter describes a platform that works in this direction.

Chapter 7

IMPACT ASSESSMENT IN URBAN FLOODS

As demonstrated in the previous chapters, in the immediate aftermath of a crisis, particularly in the first 12-24 hours, mining for ground information is of the utmost importance. For this reason, we developed a scalable, multimodal, and multilingual platform to streamline the automated processing of messages and images in near real-time. We named it the Social Media for Disaster Risk Management (SMDRM) platform.

In this chapter, we discuss the structure of the platform, we describe how we developed a model for impact assessment annotation of text, and finally, we explain how we intend to use it to improve performances of the software implemented for EFAS described in Chapter 4

The data are *collected* using keywords and locations based on daily forecasts from the early warnings systems or triggered manually in case of earthquakes or not-forecasted events. Then, the text is automatically *annotated* with multilingual classifiers trained in 12 languages and extended with multilingual embeddings. Simultaneously, a multi-class convolutional neural network labels relevant images for floods, storms, earthquakes, and fires [86]. Finally, messages are *geocoded* with a two-step algorithm; location candidates are selected using a multilingual named-entity recognition tool and then searched on available gazetteers. After the platform processing, relevant information can be aggregated in spatial (administrative areas) and temporal (daily) units.

SMDRM could offer timely, valuable information to reduce uncertainty and

provide added-value information such as reports or descriptions of the situation on the ground. The platform can help researchers to access data to complement those extracted from traditional sensors or EO. The platform can adapt to cope with surges in workload as it uses scalable software containers. Suppose the number of messages to be processed increases suddenly during a high-impact event. In that case, the platform can use more containers to annotate them. SMDRM code is released as an open-source platform.[1] Its modules can be easily extended and adapted.

In the remainder of the chapter, we describe the motivations for the platform development and the platform's architecture. We then describe the models that can be used within SMDRM, and one operational implementation.

7.1. Introduction

The usage of information from social media during emergencies has been one of the driving applications for research on the real-time processing of social media messages. Over a decade, research has sought to extract, categorize, and visualize relevant information for emergency management.

Floods have attracted significant attention for researchers, who have, for instance, attempted to determine flood extents using social media information, with some success. The uncertainty in the geolocation of messages has been reported as the main contributor to inaccuracies [8].

On average, the minimum time needed by emergency services such as the Copernicus CEMS Rapid Mapping (RM) service to provide crisis information after an activation request by an authorized user[2] is 24 hours [103]. Furthermore, due to the technical issues with densely built-up areas, remote sensing analysis is of limited use in urban areas. These areas are commonly not analyzed and left out of the produced maps.

Previous research [86] shows that social media can provide a good overview of impacted infrastructures and provide situational awareness within a few hours. Social media postings immediately after the event have a higher probability of being relevant to the event's detection and damage assessment process, and may contain less noise than later messages. This, according to practitioners, is helpful

[1] `https://github.com/ec-jrc/smdrm`

[2] EU Member States, EU Civil Protection Mechanism, the EC's Directorates-General and EU Agencies, the European External Action Service , as well as international Humanitarian Aid organizations

to crisis managers while waiting for EO products such as the ones by Copernicus Mapping.[3]

7.2. Platform Description

Concepts

A *data point* is a dictionary, typically represented in JSON format, composed of a specific set of fields described in Table 7.1.

Field	Description
id	Unique identifier
created_at	The date and time at which the data point is created
text	The textual information to be annotated and/or geo located

Table 7.1: Mandatory fields for processing data points

To **annotate** is assigning a probability score to a data point's 'text' field. This is a float number between 0 and 1, representing the likelihood that the textual information in the 'text' field is of a specified category.

A Directed Acyclic Graph (DAG) represents a *workflow* of coded instructions, which we represent within the Airflow framework.[4] A DAG specifies the workflow as a set of repeatable coded rules, including dependencies between tasks, the order to execute them, and other instructions required to run a data pipeline.

A *task* is the smallest component of a pipeline. Each task must produce the same result every time it is executed on a defined dataset. It executes a specific logic, be it fetching data, running analysis, triggering other systems, or more.

SMDRM is a Python-based data pipeline application for processing social media data points. The goal of SMDRM is to provide an enriched version of the input data shown in Table 7.2 that can be further analyzed and visualized.

[3] https://emergency.copernicus.eu/mapping/ems/cems-week-2021-conclusions-community-insights-and-service-evolutions

[4] https://airflow.apache.org/

Scalability Requirements

SMDRM application is Docker Compose[5] based. A running Docker daemon and docker-compose software are required. Considering a minimal configuration intended to run on a single machine, the workstation minimum requirements are:

- 8 CPUs
- 12 GB free memory
- 10 GB free disk storage
- access to public docker registry

Suppose multiple servers are available, or SMDRM is deployed in a production environment. In that case, we recommend setting up an orchestrated solution that runs on several machines. In that case, Docker Swarm[6] may be the easiest way, as it is configurable via settings files.

Architecture
The main components of SMDRM are:

- Docker - ensures consistency, reproducibility, and portability across Operating Systems.
- Annotators - annotate disaster types and impacts, and writes information in datapoints.
- Geocoder - extract place names from text, looks for candidates in gazetteer and and writes information in datapoints.
- Apache Airflow - authors, schedules, and monitors workflows as Directed Acyclic Graphs (DAGs) of tasks in an automated, and distributed manner.

The expected format of the input data is a zipfile archive. The zipfile should contain at least one Newline Delimited JSON (NDJSON) file. The NDJSON files must be located in the root of the zipfile archive. There must be a datapoint for each new line in the files compressed in the zipfile archive. The datapoint, including all required fields, can also be wrapped inside a 'tweet' field. This is a template typically applied to keep the original record when data is transformed.

[5] `https://www.docker.com/`
[6] `https://docs.docker.com/engine/swarm/`

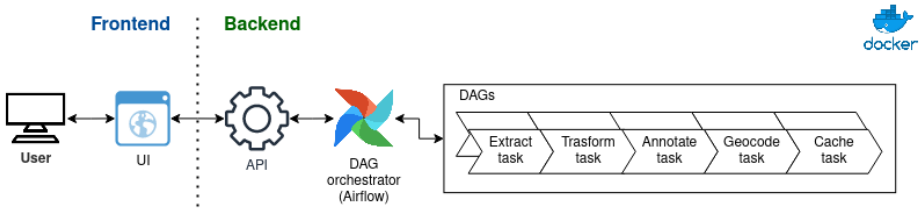


Figure 7.1: SMDRM architecture

.

7.3. Filtering, Impact Assessment, and Geocoding

Filtering (e.g., floods)

We use a supervised binary classification setup. The positive class comprised all messages indicating that a specific type of event (e.g., flood, wildfire, earthquake) is happening or is about to happen. Our work does not require semantic resources. We only leverage pre-trained encoders for multilingual modeling. They can be used to perform multilingual classification using embeddings that are aligned across languages[13]. We can use labeled data in a set of available languages to bootstrap a binary classifier for a new language for which no labels are available. In our case, we labeled data as relevant (label=1) or not(label=0) to floods or to water-related events (tropical cyclones). After several tests[51], we decided to use LASER (Language-Agnostic Sentence Representations), released by Facebook[7], as a pre-trained language representation in multiple languages. LASER acts as the encoder in our model, as it provides the embeddings for the input sentences. So, we built a classifier network for our decoder in the model to classify the sentence as relevant or not to a specific event. Specifically, we trained a sequential model with two dense layers to minimize binary cross-entropy measure as loss function using the Adam algorithm as optimizer.

Categorization (impact assessment)

We aim to make the mining of social media messages useful for practical monitoring of urban events, building upon previous work in the Joint Research Centre (JRC) Unit around Social Media Flood Monitoring.[8] We noticed soon in our

[7] shorturl.at/dwRXZ
[8] shorturl.at/bnNVX

Field	Description
annotation	Annotation scores placeholder.
place	Geographic attribures placeholder.
place.candidates	place candidates returned by NER.
place.meta	Metadata of place candidates matched against gazetteer.
place.meta.city	Name of the city.
place.meta.country	Name of the Country.
place.meta.countrycode	Alpha-3 code ISO for Country.
place.meta.latitude	Latitude of the place candidate matched against gazetteer.
place.meta.longitude	Longitude of the place candidate matched against gazetteer.
place.meta.region_id	The region identifier.
place.meta.region	The region name.
text_clean	Normalized textual information

Table 7.2: Fields added during data points processing

tests how the impacts generated from a water-related event are not different than any other disaster (i.e. injuries, evacuation, damages, services disruption) To obtain a high-quality training dataset, we performed a two-level annotation of impacts-related messages. The input to this annotation were tweets obtained during several flood or storm periods in the two cities for the pilot study. The annotators were the European VOST (Virtual Operations Support Team)[9] volunteers that process digital data for emergencies, usually composed of former or current members of various emergency response services.

Level 1: Impact / No impact The first level of annotation includes determining if a message describes an impact. In the instructions, we use the phrase 'negative impact' to avoid ambiguities in this regard and mention different types of impacts that can happen. However, We do not ask, for annotators to categorize messages based on those other types until the level 2 categorization is done.

Level 2: Type of impact The second level of annotation was focused on messages for which the level 1 annotation indicated they have an impact. Ac-

[9] `https://vosteurope.org/`

cording to our observations in the data, we considered various types of impacts that are common in urban events. First, we consider effects on specific individuals, such as people injured, missing, or displaced. Second, we split what is commonly referred to as the 'infrastructure and utilities' category into 'infrastructure damage' and 'service disruption'.

Geocoding

Depending on the aim of the application, the platform can sustain two levels of geocoding:

Regional level In our integration with EU-wide or Worldwide monitoring systems, relevant messages are mapped to NUTS-2 areas (Nomenclature of Territorial Units for Statistics, Level 2). Geocoding deals with messages that do not include explicit geographical coordinates but mention a place name such as a landmark or city. SMDRM uses a Named Entity Recognition tagger to obtain possible locations in a text considering the syntax of the message. It then uses a gazetteer in an extensive database of place names with their corresponding geographical coordinates. Finally, it uses a series of heuristics to infer the correct country and correct gazetteer entry for those places. We use a library named DeepPavlov[10] for place names identification which extracts places candidates from a piece of text. Their coordinates and structured geographic information are then searched within a list of administrative areas and cities. Messages are aggregated at the level of an event but also at the level of each administrative area.

Urban level For the application of understanding urban flood impact, messages need to be geocoded at a level of granularity that is useful for emergency responders, which in this case needs to go into an intra-urban scale, i.e., they should refer to specific areas of a city which are affected by flooding events. Only a tiny fraction of the social media messages are geotagged precisely. For instance, only 1%-5% of the tweets are geotagged within urban areas[5]. We used an innovative approach to achieve this goal. We focus on elements at risk of the infrastructure (such as a hospital, a factory, a school, or a stadium, among others), assuming that if a significant impact happens in such an infrastructure, at least some messages will mention the infrastructure by its name. However, this requires the creation of a customized gazetteer for infrastructure, which in turn requires an extensive database containing infrastructure elements' names.

[10] `http://docs.deeppavlov.ai/en/master/features/models/ner.html`

The database of known locations of infrastructure elements used in the project was built on data from OpenStreetMap (OSM).[11] Infrastructure objects in OSM are identified through 'tags' that could relate to classes defined by the Sendai Framework indicators for Disaster Risk Reduction.[12]

Figure 7.2 shows classified tweets aggregated by impact location and facilities. A manual analysis of the contents of the tweets has proved that many of the messages classified and geolocated seemed to indeed refer to impacts, some of them containing spatial references to entities which could be located close to the coordinates which were attributed to the tweets by the gazetteer.

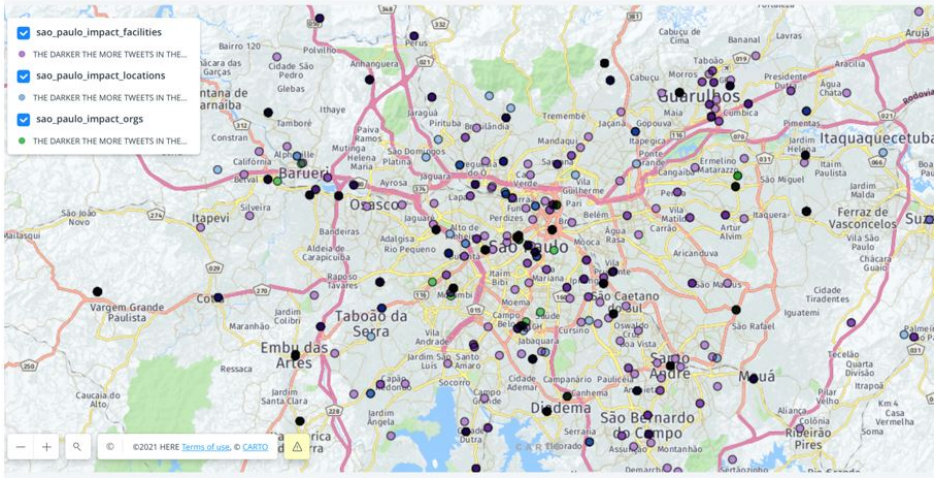


Figure 7.2: Classified tweets aggregated by impact location and facilities.

7.4. SMFR, an Instance of SMDRM

Social Media Flood Risk (SMFR) is a platform to monitor specific flood events on social media (currently, only Twitter). The system, built from the experience described in Chapter 4 is intended to work as a complementary monitoring service for existing early risk alert systems. The first release of this experimental project is tailored to work with EFAS, but in future releases, the 'topic' (floods, forest fires,etc) and the primary alert system will be configurable. For the deployment of SMFR we created a distinct DAG, and tasks tailored to the Twitter

[11] https://openstreetmap.org
[12] https://bit.ly/3OFtnFt

data structure. Each task iterates over batches of datapoints and applies a certain logic. The tasks are:

Extract: enforces the SMDRM data structure onto each datapoint in a given dataset.

Transform: applies text normalization for the annotate task, and place candidate extraction via DeepPavlov NER model for the geocode task.

Annotate: annotate data through a multilingual model for two binary classification of tweets: (i) flood relevance, expressed by a float value ranging from 0 to 1, and (ii) impact relevance, expressed by a float value ranging from 0 to 1.

Geocode: matches the place candidates extracted at transform task against the Global Places gazetteer in the case of flood-relevance , and against local gazetteers in the case of urban floods.[13]

Cache: saves processed datapoints from previous tasks into an Elastic Stack instance (Elasticsearch+Kibana) to enable data exploration/visualization in dashboard style.

Finally, we created an additional DAG, and tasks specific to the creation of products sourcing EFAS interfaces. The DAG produces GeoJSON products representing areas and risk grade, most relevant tweets per reported area, and trends per day. These files are disseminated to a list of map servers. If an area presents less than ten highly-relevant tweets (relevance $>$ 0.8), the associated region is Gray. A region is Orange if the ratio between medium-relevant tweets (0.2$<$relevance$<$0.8) and highly-relevant tweets is between 5 to 1 and 9 to 1. Region is Red if the ratio exceeds 1 to 9. After deduplication, the 5 most relevant tweets are selected and presented for each area. A product is visible in Figure 7.3. Currently the tweets annotated for "impact assessment" are not made visible as a selection of the most representative messages are analyzed by an analytical team of crisis responders.

7.5. Conclusions

Together with the image-classification model, the platform tackles *RQ1* about complementing flood risk information. The impact annotator extends the toolbox in our provision to answer *RQ5 "Is it possible to dynamically define the risk and the impact of a flood in a densely inhabited area at high resolution?"*.

One ongoing development is the implementation of a Representational State Transfer (REST) Application Programming Interface (API). This represents the

[13] `https://nominatim.openstreetmap.org`

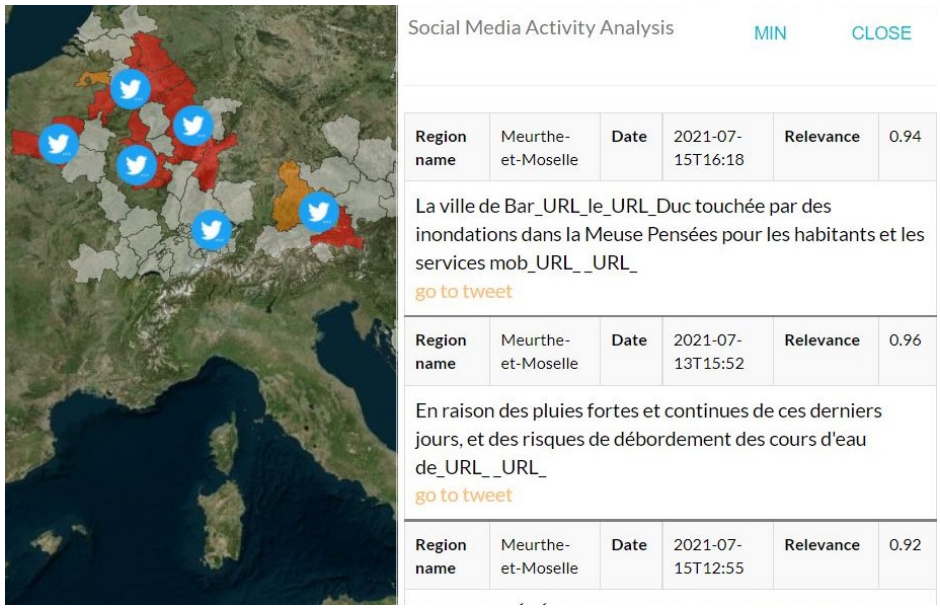


Figure 7.3: Screenshot from SMFR.

entry point where the end-user can leverage the functionalities of the SMDRM platform via HTTP requests. Components such as classification models for annotation or DAG runs can be executed through dedicated API resources. Another challenge that we are tackling is improving geocoding of messages collected during urban events. We are trying to detect partial matches between facilities and mentions in the text. As mentioned in Section 7.1 we trained a neural network for *images classification* for 5 types of disaster using the EfficientNet model structure. Image classification is currently performed by interacting with the platform asynchronously, not supervised by the Airflow component but by running offline scripts. One work in progress is to develop a module for handling media processing.

The platform presented allows to process thousands of messages at the same time, overcoming the issue of data collection and data filtering. We think that an important step forward would be to engage specific users that could evaluate our research findings and help steering our work direction. We will details our ideas in the following chapter.

Part IV

Moving Forward

Chapter 8

PRACTITIONERS' VIEW

While social media has radically reshaped many industries, how to make this abundant source of information and online behavior more accessible and usable for crisis management has remained a subject of debate – as remarked upon in a recent retrospective on crisis informatics trends and technology [80]. Part of this debate stems from possible disconnects between practitioners' needs and researchers' agendas, as practitioners believe social media has value for their domains, but key concerns that have inhibited these benefits remain unaddressed. The E1 Unit operating within the 'Space, Security and Migration' directorate of the JRC, has sought to bridge these two perspectives by organizing a workshop on 'Social Media for Disaster Risk Management: Researchers Meet Practitioners'. The workshop consisted of five panel sessions conducted online on November 30th and December 1st, 2020, and attended by 70 people.[1] Grounding this workshop in a set of emergency-related case studies, replete with numerous datasets of various sources and modalities, the workshop encouraged practitioners to share their perspectives with researchers about key aspects in which social media may provide utility in these events. Researchers shared their ideas on key technical and scientific challenges. Following this cross-disciplinary exchange, this chapter summarizes and highlights some of the main areas identified by practitioners, recognizes the perceived barriers within them, and describes the gaps between practitioner requirements and the research that need to be conducted to overcome these barriers. In making this summary available, we work towards

[1] https://publications.jrc.ec.europa.eu/repository/handle/JRC124963

establishing a common ground between practitioners and researchers for discussing methods for channeling social media information in the most fruitful way. This approach could ultimately move the field ahead by orienting the significant research effort in this space toward solving fundamental issues that have inhibited practitioners in their use of social media in crisis management.

To organize this summary, we first layout the benefits and values practitioners perceive in using social media for crisis management. We then describe the main concerns practitioners indicate that have possibly inhibited realization of these benefits and the research related to them. These issues are not new to researchers and to a large extent they have been mentioned elsewhere (e.g. [80]). Key concerns we highlight include:

- preventing negative consequences from misinformation and disinformation;
- validating social media information; and
- presenting crisis-relevant information from social media in a useful manner.

8.1. Practitioners' Perceptions of the Value of Social Media

Over the workshop, practitioners identified four areas in which they saw social media providing significant value for crisis management:

- Real-time and ongoing situational awareness;
- rapid insights in the immediate aftermath of disaster;
- integration with heterogeneous, multimodal data; and
- flexible value across the crisis management cycle.

Real-time and ongoing situational awareness. Social media's increasingly ubiquitous and global presence, particularly during disasters and emergencies, has the potential to empower decision-makers in crisis management by providing access to timely and relevant information. While, for instance, earth observation data may be delayed while satellites wait for necessary orbital positioning, or

forecasts may be inaccurate in time and space due to lack of sufficient observational data, user-generated information in social media is available 24/7/365 in real-time, all over the globe. This data can supplement risk assessment improving timeliness and efficiency of satellite based emergency mapping [16]. Recent advances in computational approaches to the massive scales of data made available through social media provide an opportunity to improve situational awareness for management and response teams.

Rapid insights in the immediate aftermath of disaster. In general, as time passes the conversation of social media tends to become more saturated with noise (e.g., jokes, political blaming, general negative sentiment, etc.) causing it to depreciate in value for practitioners. In past studies, the volume of relevant or actionable information was often considered as an indicator of an event, as it was assumed in general, that there is a spike of crisis-related messages around a particular location or topic. Recent studies [51, 14]) take a more conservative approach and integrate authoritative forecasts as well as incorporate crisis responders' feedback [77] in defining the information filtering frameworks, which can better extract and organize actionable information to inform decision makers while overcoming noise and misinformation.

An interesting case study presented during the workshop (the explosions in Beirut from August 4th, 2020) highlighted how images and videos posted by witnesses early after a disaster, can be valuable sources when no other visual depictions of a disaster exist, for instance before satellite or aerial images can become available.

The Beirut explosion is of particular note as the first image reporting the event appeared only 1 minute after the official time of the explosions (18:07 local time). This post was followed after 13 minutes by a second one showing damages at 1.5 km distance, and a third one 19 minutes later, showing the first rescuers operating in the field. In a short time, more and more images and videos appeared, providing a good understanding of the situation in the surrounding of the blast location. The images published after the first posts related to the occurrence of the event, carry the highest potential for impact assessment in terms of magnitude and location. Indeed, the average probability of images of being classified as not-relevant to the event confirms the timeline of social media activity when users, during the first hours, are mostly posting visual reports (witnesses) while later, they are joined by messages of solidarity [81, 86]

According to practitioners, beyond contributing to situational awareness, so-

cial media could be helpful in detecting sub-events within a crisis and monitoring how a crisis unfolds. Other use-cases range from performing damage assessment [8], to gathering insights about the compliance with measures and recommendations issued by authorities, and feedback about the impacts of these actions.

Integration with heterogeneous, multimodal data. As evidenced by, among others, the Beirut explosion in August 2020, the analysis of crisis-related images and videos – in addition to the standard text-analysis processes – can help humanitarian response organizations to improve decision-making and prioritize tasks. Annotated imagery and the ability to extract and unify semantic and visual features respectively from a social media post's text and images can facilitate detection and damage assessment of a crisis in new ways [38]. For example, leveraging the growing popularity of visual media from disparate sources like Instagram, YouTube and TikTok, can support detecting and assessing severity and damage from natural crises like floods, fires, landslides, earthquakes; man-made crises like industrial accident, fallout from conflict; as well as severity assessment of the damage to infrastructure and the impact on the population.

While multimodal data such as images and video are becoming more popular, so too are the needs to integrate these various kind of data into unified presentation layers. Crisis management could be improved, for example, through careful integration of this social data with other technology-based data including geospatial analytics and sensor technologies. It would require software system designed for decision support that can handle heterogeneous data from an array of platforms for social networking, media sharing, and community-driven navigation, among others, with authoritative information from radars, satellites, sensors, and other sources. This would require technical advancement to enable the capability of representing and integrating these sources in a way that is relevant for decision models of the information systems to support emergency managers.

Flexible value across the crisis management cycle. The value of social media information varies with each phase of the crisis management cycle. For example, during the preparedness phase, static social media messages from past crises can be studied to learn what type of relevant user-generated content could be anticipated in future events with similar crisis types. During the response phase, social media messages can be used to identify the extent of an event [66] or for immediate damage assessment. Social media could also be used as part of an iterative procedure for assessing efforts made by practitioners in the recovery

phase. During the mitigation phase, agencies can perform corrective measures responding to and recovering from future crises based upon lessons learned from the public social media messages communicated during past crises. For example, if evacuation orders were reported by the public through social media as 'confusing', agencies can develop clearer evacuation communication strategies in anticipation of a similar crisis in the future [87].

8.2. Barriers to Leveraging Social Media in Crisis Management.

Despite its perceived value, several key concerns have reduced trust in – and therefore adoption of – social media analysis and related technology for crisis management.

Information overload and uncertainty in automated filtering. Information overload is a critical concern when dealing with social media data. Practitioners accustomed to one-directional dissemination of information to the public are now exposed to vast amounts of data originating from the public, which precedes formal communications and exposes practitioners to overwhelming volumes of information. Of the many social media messages available during a crisis, only a very small portion of it are valuable for emergency management [68, 75, 62].

Beneficial messages have some actionable qualities, for instance, such as a clear location, as well as understandable and detailed information about a situation. Most social media messages do not fulfill this criterion. For instance, a large volume of COVID-19 tweets contain political discussions that are only tangentially valuable for crisis risk management. Therefore, it becomes essential to use reliable implementations of well-designed technologies to filter, prioritize, and organize relevant data from social media sources for decision makers. Further, these technologies need to perform in such a way that decision-making processes can be improved within a timeframe that is expected for each crisis management phase, especially during the time-sensitive response phase.

Technologies for processing social media messages, including artificial intelligence, machine learning and NLP methods, have been deployed to try to filter out irrelevant messages from social media streams. These technologies are valuable but are limited in what they can achieve. For example, it is not always clear for a person whether a message is useful or not for someone, or whether it may be useful in the future or for whom; if these assessments are difficult for

humans, automatic methods trained on these human judgement are likely to encounter difficulty. Moreover, to obtain accurate models, many algorithms need event-specific human-labelled data that may not be available in the early hours of an emergency or disaster. In addition, such filtering or categorization processes need to be clarified to overcome the concerns related to the explainability and interpretability of the automated process.

Different people and communities that must respond to disasters can be served differently by social media information, depending on its source, the type of disaster, and the timing of the information. In general, the less time passes from the moment in which a social media message is posted to the moment in which it arrives to an officer or responder, the better.

Misinformation and its consequences. Misinformation (unintentional) and disinformation (intentional) may have huge consequences with social media, including loss of life or property. There is a great differential in responsibility between users of social media platforms posting this information, and officials who may disseminate a wrong or misleading message. There are always questions of authenticity around messages posted by users of social media platforms. The public is aware of these questions, but at the same time, expects that social media channels are monitored by authorities. Sometimes the actual harm from misinformation or disinformation can be small or clearly avoidable, such as fabricated images showing sharks swimming in a flooded highway,[2] but in other cases the consequences can be large, such as false accusations to individuals.[3] Although the misleading or malicious content can be a small fraction of messages, if not removed or somehow flagged, their impact can be larger; also the fact that 'bad' content is a small fraction of the total does not mean that 'good' information abounds. Eventually, social media has some capacity for self-correction, especially in the immediate aftermath of an event, but misinformation can be extremely persistent and mislead the public as time goes by. For instance, vaccine hesitation has been, to some extent, a persistent message partially amplified by social media. Making matters worse for practitioners, traditional media outlets sometimes share the misinformation and disinformation. When stories are not thoroughly fact-checked but instead are prematurely disseminated to the public, the consequences of amplifying false or misleading information may harm a

[2]McKenzie Sadeghi: 'Fact check: Photo of shark on a flooded highway is faked'. USA Today, August 2020.

[3]'Reddit apologises for online Boston witch hunt.' BBC News, April 2013.

crisis management agency's response and recovery efforts. Time and resources need to be diverted from crisis response tasks to address false or misleading information.

Validation and verification. Practitioners want to avoid the risk of credibility loss resulting from communicating false information. To avoid this, it is imperative they check information for accuracy prior to dissemination. Also, if practitioners are to make decisions based on social media messages, the selected content must be clear, accurate and trustworthy. Validating and verifying information prior to taking action also reduces the risk of inaccurate or sub-optimal allocation of resources. False alarms and misleading or outdated information are an everyday challenge for emergency management professionals. For instance, fire fighters might mobilize to attend alerts that end up being false alarms. Practitioners are trained to validate and verify the information they receive, and one main component of this validation is the consultation with a succession of entities, and the integration of independent sources of information. In this regard, technologies that can collect and integrate heterogeneous data from various sources would be extremely valuable, particularly if they can integrate social media as a massive and dynamic source, with authoritative data. The processing of social media could be substantially improved if developed workflows can provide both timely and validated information. Efficient workflows often involve collaboration between human and automated elements. In this collaboration, the automated part would require transparency and understandable mechanisms that make them more trustworthy and easy to use by their human counterparts. As photos and videos become more prominent elements within messages, they can become more valuable as a means to aid in verifying crisis information. Recent work is making headway in this area, with new datasets of crisis-related images from social media are now available for researchers (crisisMMD).[4]) As a consequence, automated techniques to detect old, edited, or fabricated images are as important as methods for validating and verifying textual content.

Formatting information. A social media messages stream must be properly formatted before being provided to practitioners, to increase the potential for its use in decision-making. A key element of this formatting is the inclusion of geospatial information. Messages must be associated with places, regions, sites, or roads/routes/paths. In many cases, this association needs to offer high-accuracy and high-resolution, which is challenging due to the lack of an exact

[4] https://crisisnlp.qcri.org/crisismmd

geographical reference in many social media messages. Locations mentions in the text of social media posts can be extracted as a place name candidates using part-of-speech taggers and searched against gazetteers [29]. However, this approach depends on availability of data and gazetteers, and is often biased towards messages in English, limiting its applicability at a global scale. Experiences from humanitarian and collaborative mapping among other crowdsourced data collection initiatives may contribute to adding the geographical reference to social media. At the same time, more precise geolocation of messages may involve additional privacy issues when releasing fine-grained microdata. Practitioners may need data in various modalities. The first preference expressed is in a format that can be integrated within Geographical Information Systems (GIS) for visual display on a map. Secondly, in 'raw' universal textual formats such as Comma-Separated Values (CSV) or simple text files. Thirdly, in other kinds of structured or tabular form. Practitioners may also need other kinds of data according to the type of crisis. Different crisis events may require different types of reporting from social media, data formats, and levels of summarization. This highlights a need for technologies that have aspects or features that depend on the type of disaster. Finally, in many cases, multilingual data is available and need to be handled. There is, hence, also a requirement for methods that can process and collect data in multiple languages to create summaries that can be of use to the needs of different communities inhabiting a city or region.

8.3. Technical Challenges and Future Steps

Many technical challenges were named during the workshop, including the development of research, methods, and systems, to:

- Extract, transfer, and load heterogeneous data from various sources, particularly authoritative ones, and reliably integrating them into the real-time workflows, systems and tools used by practitioners for decision support.

- Automatically appraise the quality or an information source, or assist in the validation of a piece of information, be it a text message or an image or video.

- Recognizing and categorizing messages where human annotators disagree on the usefulness of the message, to generate a signal of ambiguity that can be further studied or used.

- Automatically place social media messages in time and space in an accurate manner: geocode them precisely and with high-resolution, and determine if they are timely or refer to some past or future event.
- Summarize social media messages authored in multiple languages.

As future steps, we plan to deepen our study of the gap between research and practice in this space by surveying a number of experiences of usage of social media during emergencies, and then preparing, based on that outcome, a manual of best practices that can be useful for researchers and practitioners.

Chapter 9

CONCLUSIONS AND FUTURE WORK

9.1. Summary

We collected substantial evidence to broadly assert that social media can be seen as complementary data helpful to crisis response frameworks.

Here we summarize our answers to the four research questions introduced in Chapter 1.

RQ1: Is it possible to integrate effectively social media signals with authoritative data at a pan-european level where riverine flood likelihood is estimated?
Yes, in Chapter 4, we demonstrate that using a monolingual classifier for flood relevance carried the best results in terms of accuracy. We also described a methodology to combine hydrological forecasts and automatic, immediate annotation of social media messages without translation. Such a procedure could reduce the response time extremely precious in the early stages of a flood.

RQ2: Is it possible to reliably classify social media information's relevance to floods using a 'zero-shot' transfer learning?
Yes, we demonstrated how the methodologies presented in Chapter 4 are suitable for extracting representative tweets from an event. We were able to map approximately the affected locations as demonstrated in the real event case presented.

When most messages are in a language with no labeled data, we showed how to combine the information from the classification with the known locations from the EFAS forecasts. The warm-start classifier, which involves annotating a small number of tweets in the target language, performs better than the cold-start classifier in terms of precision and recall. It often achieves an F-measure comparable to the one of the monolingual classifier. Regarding the choice of word embeddings, results suggest that the performance using GloVe or MUSE embeddings are comparable.

RQ3: Is it possible to independently identify floods from forecasts using knowledge from past events independently from hydrological forecasts?
Partially, the methodology presented in Chapter 5 can capture all types of flooding (e.g., coastal, flash flood, pluvial) with the same trained model. We observed that our model indicates flood activity (high recall) in many cases but with low precision. About one in three of the alarms generated by the model based on social media alone will correspond to a flood.

RQ4: Is it possible to dynamically define the risk and the impact of a flood in a densely inhabited area at high resolution?
In Chapter 6, we demonstrated how the accuracy of a flood extent we obtain determining whether a cell is flooded or not, using a detailed, country-specific map, arrives at 86.9%. In contrast, using maps that are freely available for the entire world grants an accuracy of 61.3% at the 10 m resolution and 65.8% at the 30 m resolution. Our experiment answers part of our research question in the affirmative, using both local DEM data and the freely available Copernicus DEM. In Chapter 7, we described a model for impact annotation. When combined with an image classifier within a scalable platform, we have a set of toolboxes for determining the development of impacts in regional and urban areas.

9.2. Future Directions

More impacts assessment
As described in Chapter 6 and Chapter 7, providing timely risk assessment at the highest possible resolution is essential. Future research should assess the feasibility of integrating information derived from social media, authoritative data (numerical models, sensors, and remote sensing), and socio-economic data. This

fusion could improve disaster risk management capacity, especially in urban areas, and study the correlation between social media activity and flood impacts. The research can explore the latest technologies to classify the information derived from, but not limited to, multilingual social media, news, traffic data, and publicly available data in real-time. An additional research effort could analyze the information derived from events that are reported in the ground truth database but are not found in Wikipedia.

Mapping

Limited human resources compound the issues mentioned above within emergency services with pre-assigned responsibilities. The volume of social media data generated during disasters risks cognitive overload for the human resources who perform multiple tasks. Since the content from social media platforms can be multimodal, this requires specific techniques to collect, store, integrate, and analyze the information. We think future research should improve the timing and precision of identifying locations impacted by a natural hazard. Leveraging the methodology described in Chapter 6 and Chapter 7, images could be filtered automatically based on the hypothesis that some images' features could facilitate their geocoding for a specific disaster. Such a line of research should also focus on the multimodal information because, as presented in Chapter 3, text and images combined could help improve human geocoding and machine annotation.

Engaging practitioners

In Chapter 8, we described the barriers to adopting social media analysis in crisis response. They include a variety of information quality and processing issues. Social media data's reliability can sometimes lead to negative consequences, such as emergency services officers exerting time and energy to sift through misinformation and disinformation; careful pre-processing of validated and relevant data is necessary. Social media data resolution can be inconsistent; sometimes, information comes with the geolocation metadata at the fine-grained level if users enable it. Others come at an aggregated, coarse-grained level.
Furthermore, there are multiple social media platforms. Thus, these data sources have the characteristics of heterogeneity in terms of format, metadata, and structure. Lastly, the content from social media platforms can be multimodal, requiring specific techniques to collect, store, integrate, and analyze the information. We think that a future line of research is identifying a common framework for

exchanging information between researchers and practitioners. It should study data formats, software modules, and procedures to make information processable by tools deployed across emergency centers without disrupting emergency operations. This novel approach should facilitate an agile exchange of tasks/products between data analysts and crisis responders, providing a non-invasive capability to distill the essential information without overtime or overloading crisis rooms with too much data. Given the focus on making a tangible impact on the practice, we think a co-designing approach is needed to ensure practitioners' participation.

Ultimately, We think we can say that the scope of the research must be extended to other types of events. To scale the current methodology, researchers need to identify adequate global ground truth information for the specific type of event and a classifier for it.

Bibliography

[1] F. Alam, T. Alam, F. Ofli, and M. Imran, «Social media images classification models for real-time disaster response», 2021. arXiv: 2104.04184.

[2] F. Alam, F. Ofli, and M. Imran, «Crisismmd: Multimodal twitter datasets from natural disasters», *CoRR*, vol. abs/1805.00713, 2018. arXiv: 1805.00713. [Online]. Available: http://arxiv.org/abs/1805.00713.

[3] L. Alfieri, P. Burek, L Feyen, and G. Forzieri, «Global warming increases the frequency of river floods in Europe», *Hydrology and Earth System Sciences Discussions*, vol. 12, pp. 1119–1152, May 2015. DOI: 10.5194/hessd-12-1119-2015.

[4] L. Alfieri, V. Lorini, F. A. Hirpa, S. Harrigan, E. Zsoter, C. Prudhomme, and P. Salamon, «A global streamflow reanalysis for 1980–2018», *Journal of Hydrology X*, vol. 6, p. 100 049, 2020, ISSN: 2589-9155. DOI: https://doi.org/10.1016/j.hydroa.2019.100049.

[5] S. de Andrade, J. De Albuquerque, R. Westerholt, C. Restrepo-Estrada, C. Morales, E. Mendiondo, and A. Delbem, «The effect of intra-urban mobility flows on the spatial heterogeneity of social media activity: Investigating the response to rainfall events», *Int. J. of Geo. Inf. Sci.*, Aug. 2021. DOI: 10.1080/13658816.2021.1957898.

[6] M. Artetxe and H. Schwenk, «Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond», *CoRR*, vol. abs/1812.10464, 2018. arXiv: 1812.10464. [Online]. Available: http://arxiv.org/abs/1812.10464.

[7] F. Botta, H. S. Moat, and T. Preis, «Measuring the size of a crowd using instagram», *Environment and Planning B: Urban Analytics and City Science*, 2019. DOI: 10.1177/2399808319841615.

[8] T. Brouwer, D. Eilander, A. van Loenen, M. J. Booij, K. M. Wijnberg, J. S. Verkade, and J. Wagemaker, «Probabilistic flood extent estimates from social media flood observations», *Natural Hazards and Earth System Sciences*, vol. 17, no. 5, pp. 735–747, 2017. DOI: 10.5194/nhess-17-735-2017.

[9] E. S. Callahan and S. C. Herring, «Cultural bias in wikipedia content on famous persons», *Journal of the American Society for Information Science and Technology*, vol. 62, no. 10, pp. 1899–1915, 2011. DOI: 10.1002/asi.21577.

[10] C. Castillo, *Big Crisis Data*. Cambridge University Press, Jul. 2016. DOI: 10.1017/CBO9781316476840.

[11] L. Cavaleri *et al.*, «The 2019 flooding of venice and its implications for future predictions», *Oceanography*, vol. 1, Mar. 2020. DOI: 10.5670/oceanog.2020.105.

[12] D. Chicco and G. Jurman, «The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation», *BMC Genomics*, vol. 21, 2020, ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7.

[13] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, «Word Translation Without Parallel Data», en, *arXiv:1710.04087 [cs]*, Oct. 2017, arXiv: 1710.04087. [Online]. Available: http://arxiv.org/abs/1710.04087 (visited on 10/25/2018).

[14] J. A. de Bruijn, H. de Moel, A. H. Weerts, M. C. de Ruiter, E. Basar, D. Eilander, and J. C. Aerts, «Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network», *Computers & Geosciences*, vol. 140, p. 104 485, 2020, ISSN: 0098-3004. DOI: https://doi.org/10.1016/j.cageo.2020.104485.

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, «BERT: pre-training of deep bidirectional transformers for language understanding», *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[16] F. Dottori, M. Kalas, P. Salamon, A. Bianchi, L. Alfieri, and L. Feyen, «An operational procedure for rapid flood risk assessment in Europe», English, *Natural Hazards and Earth System Sciences*, vol. 17, no. 7, pp. 1111–1126, Jul. 2017, ISSN: 1561-8633. DOI: https://doi.org/10.5194/nhess-17-1111-2017. (visited on 02/13/2019).

[17] F. Dottori, P. Salamon, A. Bianchi, L. Alfieri, F. A. Hirpa, and L. Feyen, «Development and evaluation of a framework for global flood hazard mapping», *Advances in Water Resources*, vol. 94, pp. 87–102, Aug. 2016, ISSN: 0309-1708. DOI: 10.1016/j.advwatres.2016.05.002. (visited on 10/25/2018).

[18] F. Dottori *et al.*, «Satellites, tweets, forecasts: The future of flood disaster management?», in *EGU General Assembly Conference Abstracts*, 2017.

[19] F. Dottori *et al.*, «Increased human and economic losses from river flooding with anthropogenic warming», *Nature Climate Change*, vol. 8, no. 9, pp. 781–786, 2018, ISSN: 1758-6798. DOI: 10.1038/s41558-018-0257-z.

[20] EC, *Nomenclature of territorial units for statistics*, 2016. [Online]. Available: https://ec.europa.eu/eurostat/web/nuts/background/.

[21] G. Erkan and D. R. Radev, «Lexrank: Graph-based lexical centrality as salience in text summarization», *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[22] M. Ferron and P. Massa, «Collective memory building in wikipedia: The case of north african uprisings», in *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, ACM, 2011, pp. 114–123.

[23] G. Formetta and L. Feyen, «Empirical evidence of declining global vulnerability to climate-related hazards», *Global Environmental Change*, vol. 57, May 2019. DOI: 10.1016/j.gloenvcha.2019.05.004.

[24] J. Galtung and M. Ruge, «The structure of foreign news», *Journal of Peace Research - J PEACE RES*, vol. 2, pp. 64–90, Mar. 1965. DOI: 10.1177/002234336500200104.

[25] A. Gebrehiwot, L. Hashemi-Beni, G. Thompson, P. Kordjamshidi, and T. E. Langan, «Deep convolutional neural network for flood extent mapping using unmanned aerial vehicles data», *Sensors*, vol. 19, no. 7, 2019, ISSN: 1424-8220. DOI: 10.3390/s19071486.

[26] M. Graham, B. Hogan, R. K. Straumann, and A. Medhat, «Uneven geographies of user-generated information: Patterns of increasing informational poverty», *Annals of the Association of American Geographers*, vol. 104, no. 4, pp. 746–764, 2014.

[27] M. B. Gurstein, «Open data: Empowering the empowered or effective data use for everyone?», *First Monday*, vol. 16, no. 2, 2011.

[28] T. Ha, J. Niehues, and A. H. Waibel, «Toward multilingual neural machine translation with universal encoder and decoder», 2016. arXiv: 1611.04798. [Online]. Available: http://arxiv.org/abs/1611.04798.

[29] A. Halterman, «Mordecai: Full text geoparsing and event geocoding», *The Journal of Open Source Software*, vol. 2, no. 9, 2017. DOI: 10.21105/joss.00091.

[30] B. Hecht and D. Gergle, «Measuring self-focus bias in community-maintained knowledge repositories», in *Proceedings of the fourth international conference on Communities and technologies*, ACM, 2009, pp. 11–20.

[31] K. Hiroki, *2019 help global report on water and disasters*, 2019. [Online]. Available: http://www.wateranddisaster.org/cms310261/wp-content/uploads/2019/07/HELP-Global-Report-on-Water-and-Disasters-D9-20190607_s.pdf.

[32] F. A. Hirpa, P. Salamon, H. E. Beck, V. Lorini, L. Alfieri, E. Zsoter, and S. J. Dadson, «Calibration of the global flood awareness system (glofas) using daily streamflow data», *Journal of Hydrology*, vol. 566, pp. 595 –606, 2018, ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2018.09.052.

[33] C. Hube, «Bias in wikipedia», in *Proceedings of the 26th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, 2017, pp. 717–721, ISBN: 978-1-4503-4914-7. DOI: 10.1145/3041021.3053375.

[34] E. Hutchins, *Cognition in the Wild*. MIT press, 1995.

[35] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, «Processing social media messages in mass emergency: A survey», *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.

[36] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, «Aidr: Artificial intelligence for disaster response», in *Proceedings of the 23rd International Conference on World Wide Web*, ACM, 2014, pp. 159–162.

[37] M. Imran, P. Mitra, and C. Castillo, «Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages», in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, European Language Resources Association (ELRA), 2016, pp. 1638–1643.

[38] M. Imran, F. Ofli, D. Caragea, and A. Torralba, «Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions», *Information Processing & Management*, vol. 57, no. 5, p. 102 261, 2020, ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2020.102261.

[39] T. Joachims, «Svmlight: Support vector machine», *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, vol. 19, no. 4, 1999.

[40] I. L. Johnson, Y. Lin, T. J.-J. Li, A. Hall, A. Halfaker, J. Schöning, and B. Hecht, «Not at home on the range: Peer production and the urban/rural divide», in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 2016, pp. 13–25.

[41] M. Johnson *et al.*, «Google's multilingual neural machine translation system: Enabling zero-shot translation», *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017. DOI: 10.1162/tacl_a_00065.

[42] B. Keegan, «A history of newswork on wikipedia», in *Proceedings of the 9th international symposium on open collaboration*, ACM, 2013, p. 7.

[43] B. Keegan, D. Gergle, and N. Contractor, «Hot off the wiki: Dynamics, practices, and structures in wikipedia's coverage of the tōhoku catastrophes», in *Proceedings of the 7th international symposium on Wikis and open collaboration*, ACM, 2011, pp. 105–113.

[44] ——, «Staying in the loop: Structure and dynamics of wikipedia's breaking news collaborations», in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, ACM, 2012, p. 1.

[45] P. Khare, G. Burel, D. Maynard, and H. Alani, «Cross-Lingual Classification of Crisis Data», en, in *The Semantic Web – ISWC 2018*, D. Vrandečić *et al.*, Eds., vol. 11136, Cham: Springer International Publishing, 2018, pp. 617–633. DOI: 10.1007/978-3-030-00671-6_36. (visited on 10/25/2018).

[46] H. Li, X. Li, D. Caragea, and C. Caragea, «Comparison of word embeddings and sentence encodings as generalized representations for crisis tweet classification tasks», 2018.

[47] R. Liao, *Text classification, part i - convolutional networks*, 2016. [Online]. Available: https://richliao.github.io/supervised/classification/2016/11/26/textclassifier-convolutional/.

[48] A. C. E. Lima and L. Nunes de Castro, «A multi-label, semi-supervised classification approach applied to personality prediction in social media», *Neural Networks*, vol. 58, pp. 122 –130, 2014, Special Issue on "Affective Neural Networks and Cognitive Learning Systems for Big Data Analysis", ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2014.05.020.

[49] Y. N. Lin, S.-H. Yun, A. Bhardwaj, and E. M. Hill, «Urban flood detection with sentinel-1 multi-temporal synthetic aperture radar (sar) observations in a bayesian framework: A case study for hurricane matthew», *Remote Sensing*, vol. 11, no. 15, 2019, ISSN: 2072-4292. DOI: 10.3390/rs11151778.

[50] S. M. Liu and J.-H. Chen, «A multi-label classification based approach for sentiment classification», *Expert Systems with Applications*, vol. 42, no. 3, pp. 1083 –1093, 2015, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2014.08.036.

[51] V. Lorini, C. Castillo, F. Dottori, M. Kalas, D. Nappo, and P. Salamon, «Integrating social media into a pan-european flood awareness system:A multilingual approach», in *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, València, Spain, May 19-22, 2019*, ISCRAM Association, 2019. [Online]. Available: http://idl.iscram.org/files/valeriolorini/2019/1854_ValerioLorini_etal2019.pdf.

[52] V. Lorini, C. Castillo, D. Nappo, F. Dottori, and P. Salamon, «Social media alerts can improve, but not replace hydrological models for forecasting floods», in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2020, pp. 351–356. DOI: 10.1109/WIIAT50758.2020.00050.

[53] V. Lorini, E. Panizio, and C. Castillo, «Smdrm: A platform to analyze social media for disaster risk management in near real time», *Proceedings of Workshop on Social Media for Emergency Response SOMMER22*, 2022.

[54] V. Lorini, J. Rando, D. Saez-Trumper, and C. Castillo, «Uneven coverage of natural disasters in wikipedia: The case of floods», in *ISCRAM 2020 Conference Proceedings – 17th International Conference on Information Systems for Crisis Response and Management*, ISCRAM Association, 2020.

[55] V. Lorini, P. Rufolo, and C. Castillo, «Venice was flooding... one tweet at a time.», in *Proceedings of ACM Conference On Computer-Supported Cooperative Work And Social Computing 2022*, CSCW, 2022.

[56] V. Lorini *et al.*, «Social media for emergency management: Opportunities and challenges at the intersection of research and practice», in *ISCRAM 2021 Conference Proceedings – 18th International Conference on Information Systems for Crisis Response and Management*, ISCRAM Association, 2021.

[57] T. Luong, H. Pham, and C. D. Manning, «Bilingual word representations with monolingual quality in mind», in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, Association for Computational Linguistics, 2015, pp. 151–159. DOI: 10.3115/v1/W15-1521. (visited on 11/28/2018).

[58] M. Marin Ferrer, L. Vernaccini, and K. Poljansek, *Inform index for risk management: Concept and methodology, version 2017*, 2017. DOI: 10.2760/08037.

[59] D. C. Mason, S. L. Dance, and H. L. Cloke, «Floodwater detection in urban areas using sentinel-1 and worlddem data», *Journal of Applied Remote Sensing*, vol. 15, no. 3, pp. 1 –22, 2021. DOI: 10.1117/1.JRS.15.032003.

[60] D. c. Mason, S. L. Dance, S. Vetra-Carvalho, and H. L. Cloke, «Robust algorithm for detecting floodwater in urban areas using synthetic aperture radar images», *Journal of Applied Remote Sensing*, vol. 12, no. 4, pp. 1 –20, 2018. DOI: 10.1117/1.JRS.12.045011.

[61] F. McClean, R. Dawson, and C. Kilsby, «Implications of using global digital elevation models for flood risk analysis in cities», *Water Resources Research*, vol. 56, no. 10, 2020. DOI: https://doi.org/10.1029/2020WR028241.

[62] R. Mccreadie, C. Buntain, and I. Soboroff, «TREC Incident Streams: Finding Actionable Information on Social Media», in *Proceedings of the 16th International Conference on Information Systems for Crisis Response And Management*, 2019, pp. 691–705. [Online]. Available: http://cody.bunta.in/.

[63] S. E. Middleton, A. Zielinski, L. N. Tokarchuk, and X. Wang, «Social-media text mining and network analysis to support decision support for natural crisis management», presented at the ISCRAM 2013, 2013. [Online]. Available: https://eprints.soton.ac.uk/359364/.

[64] D. Moats, «Following the fukushima disaster on (and against) wikipedia: A methodological note about sts research and online platforms», *Science, Technology, & Human Values*, vol. 44, no. 6, pp. 938–964, 2019.

[65] A. Musaev, D. Wang, and C. Pu, «LITMUS: A Multi-Service Composition System for Landslide Detection», *IEEE Transactions on Services Computing*, vol. 8, no. 5, pp. 715–726, Sep. 2015, ISSN: 1939-1374. DOI: 10.1109/TSC.2014.2376558.

[66] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, «Damage assessment from social media imagery data during disasters», in *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2017, pp. 569–576.

[67] A. Nobre *et al.*, «Height above the nearest drainage – a hydrologically relevant new terrain model», *Journal of Hydrology*, vol. 404, no. 1, pp. 13–29, 2011, ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2011.03.051.

[68] A. Olteanu, S. Vieweg, and C. Castillo, «What to expect when the unexpected happens: Social media communications across crises», in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '15, ACM, 2015, pp. 994–1009, ISBN: 978-1-4503-2922-4. DOI: 10.1145/2675133.2675242.

[69] L. Palen and K. M. Anderson, «Crisis informatics—new data for extraordinary times», *Science*, vol. 353, no. 6296, pp. 224–225, 2016.

[70] S. M. Papalexiou and A. Montanari, «Global and regional increase of precipitation extremes under global warming», *Water Resources Research*, vol. 55, no. 6, pp. 4901–4914, 2019. DOI: 10.1029/2018WR024067.

[71] S. Park, J. Serrà, E. F. Martinez, and N. Oliver, «Mobinsight: A framework using semantic neighborhood features for localized interpretations of urban mobility», *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 3, Jul. 2018, ISSN: 2160-6455. DOI: 10.1145/3158433.

[72] D. Pastor-Escuredo, Y. Torres, M. Martinez, and P. J. Zufiria, «Floods impact dynamics quantified from big data sources», *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1804.09129.

[73] J. Pennington, R. Socher, and C. Manning, «GloVe: Global vectors for word representation», in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[74] B. Poblete, J. Guzmán, J. Maldonado, and F. Tobar, «Robust detection of extreme events using twitter: Worldwide earthquake monitoring», *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2551–2561, 2018.

[75] H. Purohit, C. Castillo, M. Imran, and R. Pandev, «Social-EOC: Serviceability model to rank social media requests for emergency operation centers», *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pp. 119–126, 2018. DOI: 10.1109/ASONAM.2018.8508709.

[76] H. Purohit, A. Hampton, S. Bhatt, V. L. Shalin, A. P. Sheth, and J. M. Flach, «Identifying seekers and suppliers in social media communities to support crisis coordination», *Computer Supported Cooperative Work (CSCW)*, vol. 23, no. 4, pp. 513–545, Dec. 2014, ISSN: 1573-7551. DOI: 10.1007/s10606-014-9209-y.

[77] H. Purohit and S. Peterson, «Social media mining for disaster management and community resilience», in *Big Data in Emergency Management: Exploitation Techniques for Social and Mobile Data*, Springer, 2020, pp. 93–107.

[78] Redazione ANSA, *Calabria bad weather 201810*, http://www.ansa.it/english/news/2018/10/05/mum-child-dead-as-extreme-weather-batters-south_dd387479-55cb-4249-ba69-09862501bef2.html/, 2018.

[79] C. Restrepo-Estrada, S. C. de Andrade, N. Abe, M. C. Fava, E. M. Mendiondo, and J. P. de Albuquerque, «Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring», en, *Computers & Geosciences*, vol. 111, pp. 148–158, Feb. 2018, ISSN: 00983004. DOI: 10.1016/j.cageo.2017.10.010.

[80] C. Reuter, G. Backfried, M.-A. Kaufhold, and F. Spahr, «Iscram turns 15: A trend analysis of social media papers 2004-2017», *Proceedings of the 15th International ISCRAM Conference*, no. May, pp. 1–14, 2018.

[81] Y. Rizk, H. S. Jomaa, M. Awad, and C. Castillo, «A computationally efficient multi-modal classification approach of disaster-related twitter images», in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, New York, NY, USA: Association for Computing Machinery, 2019, 2050–2059, ISBN: 9781450359337. DOI: 10.1145/3297280.3297481.

[82] B. Robinson, R. Power, and M. Cameron, «A sensitive twitter earthquake detector», in *Proceedings of the 22nd international conference on world wide web*, 2013, pp. 999–1002.

[83] J. F. Rosser, D. G. Leibovici, and M. J. Jackson, «Rapid flood inundation mapping using social media, remote sensing and topographic data», *Natural Hazards: Journal of the International Society for the Prevention and Mitigation of Natural Hazards*, vol. 87, no. 1, pp. 103–120, May 2017. DOI: 10.1007/s11069-017-2755-0.

[84] C. Rossi *et al.*, «Early detection and information extraction for weather-induced floods using social media streams», *International Journal of Disaster Risk Reduction*, vol. 30, pp. 145 –157, 2018, ISSN: 2212-4209. DOI: https://doi.org/10.1016/j.ijdrr.2018.03.002.

[85] H. Rossling, *Factfullness*. Flammarion, 2019.

[86] P. Rufolo, D. Muraro, and V. Lorini, «Social media image analysis in the immediate aftermath of the 2020 beirut blast, 2021», vol. 50, 2021, ISSN: 1831-9424. DOI: 10.2760/944555.

[87] I. Ruin, S. Shabou, S. Chardonnel, C. Lutoff, and S. Anquetin, «When driving to work becomes dangerous», in *Mobility in the Face of Extreme Hydrometeorological Events 2*, C. Lutoff and S. Durand, Eds., ISTE, 2020, pp. 91–118, ISBN: 978-1-78548-290-8. DOI: https://doi.org/10.1016/B978-1-78548-290-8.50004-3.

[88] N. Said *et al.*, «Deep learning approaches for flood classification and flood aftermath detection», *Working Notes Proceedings of the MediaEval 2018 Workshop*, vol. 2283, 2018.

[89] A. Samoilenko, F. Lemmerich, K. Weller, M. Zens, and M. Strohmaier, «Analysing timelines of national histories across wikipedia editions: A comparative computational approach», in *In Proc. of ICWSM*, 2017.

[90] R. Schiaroli, *Alluvione calabria 201810*, https://www.centrometeoitaliano.it/notizie-meteo/alluvione-calabria-2018-caduti-oltre-300-millimetri-di-pioggia-dal-2-al-5-ottobre-8-10-2018-67267/, 2018.

[91] A. Severyn and A. Moschitti, «Twitter sentiment analysis with deep convolutional neural networks», in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2015, pp. 959–962.

[92] E. Sheehan *et al.*, «Predicting economic development using geolocated wikipedia articles», in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19, ACM, 2019, pp. 2698–2706, ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330784.

[93] P. Smith *et al.*, «Chapter 11 - on the operational implementation of the european flood awareness system (efas)», in *Flood Forecasting*, T. E. Adams and T. C. Pagano, Eds., Boston: Academic Press, 2016, pp. 313–348, ISBN: 978-0-12-801884-2. DOI: https://doi.org/10.1016/B978-0-12-801884-2.00011-6.

[94] T. Steiner and R. Verborgh, «Disaster monitoring with wikipedia and online social networking sites: Structured data and linked data fragments to the rescue?», in *2015 AAAI Spring Symposium Series*, 2015. [Online]. Available: http://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10272.

[95] S. Tarquini, I. Isola, M. Favalli, M. Mazzarini F. Bisson, M. Pareschi, and E. Boschi, «Tinitaly/01: A new triangular irregular network of italy, 2007», *Ann. Geophys.*, vol. 50, no. 3, 2007. DOI: 10.4401/ag-4424.

[96] W. R. Tobler, «A computer movie simulating urban growth in the detroit region», *Economic geography*, vol. 46, no. sup1, pp. 234–240, 1970.

[97] I. Ullah, S. Khan, M. Imran, and Y.-K. Lee, «Rweetminer: Automatic identification and categorization of help requests on twitter during disasters», *Expert Systems with Applications*, vol. 176, p. 114 787, 2021, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.114787.

[98] UN Office for the Coordination of Humanitarian Affairs, «World humanitarian data and trends», 2017.

[99] UNDRR, *2019 global assessment report on disaster risk reduction*, 2019.

[100] S. Vieweg, C. Castillo, and M. Imran, «Integrating social media communications into the rapid assessment of sudden onset disasters», en, in *Social Informatics*, vol. 8851, Cham: Springer International Publishing, 2014, pp. 444–461. DOI: 10.1007/978-3-319-13734-6_32. (visited on 10/25/2018).

[101] M. Vousdoukas, L. Mentaschi, E. Voukouvalas, M. Verlaan, S. Jevrejeva, L. P. Jackson, and L. Feyen, «Global probabilistic projections of extreme sea levels show intensification of coastal flood hazard», *Nature Communications*, vol. 9, Dec. 2018. DOI: 10.1038/s41467-018-04692-w.

[102] R.-Q. Wang, H. Mao, Y. Wang, C. Rae, and W. Shaw, «Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data», *Computers & Geosciences*, vol. 111, pp. 139–147, 2018, ISSN: 0098-3004. DOI: https://doi.org/10.1016/j.cageo.2017.11.008.

[103] A. Wania, I. Joubert-Boitat, F. Dottori, M. Kalas, and P. Salamon, «Increasing timeliness of satellite-based flood mapping using early warning systems in the copernicus emergency management service», *Remote Sensing*, vol. 13, no. 11, 2021, ISSN: 2072-4292. DOI: 10.3390/rs13112114.

[104] A. Wirtz, W. Kron, P. Low, and M. Steuer, «The need for data: Natural disasters and the challenges of database management», *Natural Hazards*, vol. 70, no. 1, pp. 135–157, 2014, ISSN: 1573-0840. DOI: 10.1007/s11069-012-0312-4.

[105] D. Wu and Y. Cui, «Disaster early warning and damage assessment analysis using social media data and geo-location information», *Decision Support Systems*, vol. 111, pp. 48 –59, 2018, ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2018.04.005.

[106] H. Wu, R. F. Adler, Y. Tian, G. J. Huffman, H. Li, and J. Wang, «Real-time global flood estimation using satellite-based precipitation and a coupled land surface and routing model», *Water Resources Research*, vol. 50, no. 3, pp. 2693–2717, 2014. DOI: 10.1002/2013WR014710.

[107] L. Xiaoyan, Y. Saini, Y. Tao, A. Rui, and C. Cuizhen, «A new approach to estimating flood-affected populations by combining mobility patterns with multi-source data: A case study of wuhan, china», *International Journal of Disaster Risk Reduction*, vol. 55, Feb. 2021. DOI: 10.1016/j.ijdrr.2021.102106.

[108] K. K. Yilmaz, R. Adler, Y. Hong, J. Wang, F. Policelli, Y. Tian, and H. Pierce, «Update on nasa's real-time global flood monitoring system: Recent improvements and examples», in *EGU General Assembly Conference Abstracts*, vol. 12, 2010, p. 7798.

[109] L. Yue, L. Qinghua, and S. Jie, «Discover patterns and mobility of twitter users—a study of four us college cities», *ISPRS International Journal of Geo-Information*, vol. 6, no. 2, 2017, ISSN: 2220-9964. DOI: 10.3390/ijgi6020042.