Early Detection and Information Extraction for Weather-induced Floods using Social Media Streams

C. Rossi[a], F. S. Acerbo[b], K. Ylinen[c], I. Juga[c], P. Nurmi[c], A. Bosca[d], F. Tarasconi[d], M. Cristoforetti[e], A. Alikadic[e]

[a]*Istituto Superiore Mario Boella (ISMB), Torino, Italy*
[b]*Politecnico di Torino, Italy*
[c]*Finnish Meteorological Institute, Helsinki, Finland*
[d]*CELI Language Technology, Torino, Italy*
[e]*Fondazione Bruno Kessler (FBK), Trento, Italy*

Abstract

Today we are using an unprecedented wealth of social media platforms to generate and share information regarding a wide class of events, which include extreme meteorological conditions and natural hazards such as floods. This paper proposes an automated set of services that start from the availability of weather forecasts, including both an event detection technique and a selective information retrieval from on-line social media. The envisioned services aim to provide qualitative feedback for meteorological models, detect the occurrence of an emergency event and extract informative content that can be used to complement the situational awareness. We implement such services and evaluate them during a recent weather induced flood. Our approach could be highly beneficial for monitoring agencies and meteorological offices, who act in the early warning phase, and also for authorities and first responders, who manage the emergency response phase.

Keywords: extreme weather, flood, social media, text mining, anomaly detection, classification

1. Introduction

It is commonly acknowledged that high impact, extreme weather events occur more frequently and last longer due to climate change. During the last 35 years, the average Earth surface temperature has risen about 0.8 °C [1]. According to the Intergovernmental Panel on Climate Change (IPCC), the surface temperature is projected to rise throughout the 21st century under all assessed emission scenarios [2]. Such global warming directly affects precipitations because the water holding capacity of air increases by about 7% per degree C [3] that leads to more water vapor being retained in the atmosphere. Storms, thunderstorms, extra-tropical rains, snow, are therefore supplied with more moisture and produce more extreme precipitation events. Such events are observed to be widely occurring, even where total precipitation is decreasing, and, in combination with rapid snow melting, they increases the risk of flooding.

Given that floods are usually weather-induced, meteorological services provide local authorities with a periodical weather and flood hazard forecast that contains an encoded alert level on a predetermined set of geographical areas. The alert level is used to trigger actions according to a predefined operational procedure, which can encompass monitoring activities aimed at assessing in-field circumstances or at rapidly detecting the occurrence of the flood.

When a flood strikes, authorities and first responders can rely on satellite-based mapping (e.g., through Copernicus EMS [4]) in order to understand the extend and the impact of floods both in the response and in the post disaster phase. One of the most significant transformations in cartography over the last years has been the radical shift from static maps to live and dynamic maps. The growing volume of real-time geo-referenced data and the availability of multiple data sources are largely responsible for this shift towards real-time mapping. The data is generated all over the world both from physical sensors and from humans collecting the data. Despite highly specialized and capable emergency management systems, ordinary citizens are usually the first on the scene in an emergency or disaster, and remain long after official services have ceased. Citizens often play vital roles in helping the emergency response and the recovery of the affected individuals, and can provide valuable assistance to official agencies. People equipped with mobile devices act as a mass of multimedia sensors. This evolving network of human sensors generates a significant amount of real-time data, especially via social media platforms such as Facebook, YouTube, Flickr, and Twitter, which is the most widely used in times of crisis [5].

The use of on-line social media platforms during emergency events, coupled with the ubiquity of mobile devices capable of providing high-resolution geolocated multimedia content, offers the opportunity to exploit the generated data in order to (i) detect the occurrence of an event in real time, and

(ii) gather useful real-time on-field observations in order to improve satellite mapping and situational awareness.

However, including data from social media in emergency management processes poses several challenges, including the availability of location information, the truthfulness and accuracy of the shared information, as well as the big volume, velocity, and variety of data.

This paper assesses the feasibility to establish an automatic set of services aimed at linking weather forecasting with event detection and information extraction using social media streams. We take as case study the data generated within Twitter, before and during a recent weather-induced flood in north Italy, assessing the dynamics of the data generation process and the extraction of valuable information for the key stakeholders of emergency management: meteorological agencies, who issue weather forecasts and alerts, and first responders, who have to act in the response phase.

The paper is organized as follows. In Section 2 we review related works on extreme weather forecasting and social media analysis for emergency management, while in Section 3 we describe our case study. In Section 4 we outline the proposed solution and the components involved. Section 5 describes the methodology adopted by the different components, while in Sections 6 and 7 implementations and results are presented, respectively. Finally, conclusions and future works are outlined in Section 8.

2. Related Works

2.1. Weather Extremes: impact on society

Extreme weather conditions can cause disruption of critical infrastructures, damage to private and public assets, and even deaths. The impacts of extreme weather events on society have been recently investigated in numerous studies, e.g., EU-funded projects EWENT [1], MOWE-IT [2] and RAIN [3]. Both the EWENT and MOWE-IT projects focused on the impacts of adverse weather on the European transportation system, whereas in RAIN the focus was on four types of Critical Infrastructures (CI): roads, railways, electric power supplies and telecommunication infrastructure. The outcome of the RAIN project revealed that the most important weather phenomena having

[1] www.ewent.vtt.fi

[2] www.mowe-it.eu

[3] rain-project.eu

negative impacts on CIs are freezing precipitation, snowfall, snow loading and snow storms, windstorms and heavy precipitation causing flooding [6].

A common practice within national weather services and meteorological forecasters is to issue warnings against adverse weather events based on specific thresholds, which are relevant for a given region. The warnings typically cover a 24-hour or 48-hour time span, but many weather services also produce the so-called early warnings in the 2-5 day range. Warnings at the European level are provided by the Meteoalarm [4] service under the EUMETNET (European Meteorological Services Network) umbrella, where most European national weather services generate the original local input to the Meteoalarm framework.

Heavy precipitation events often trigger severe floods that can cause large damages. Rainfall can be highly variable with respect to duration, intensity or spatial extent. Both short-duration and heavy downpours or long-lasting and moderate rainfalls can have negative impacts. The stakeholder and weather service interviews realized within the RAIN project revealed that a universal impact-threshold value cannot be defined for heavy precipitation. The thresholds being highlighted varied between 20 mm/hour to 30 mm/hour for short-term heavy precipitation events, and from 50 mm/day up to 100 mm/day for longer-lasting rain events [7]. Instead of using fixed precipitation threshold values, another approach is to use local return values, i.e. the amount of precipitation per time unit, exceeded on average every N years (N being for example 5, 10, 50, 100 etc.) [7], [8]. This method is suitable for research purposes, whereas the use of a fixed threshold is more convenient for operational forecasting and warning procedures. Figure 1 shows the distribution of the 10-year return level for 24-hour precipitation in Europe. The highest values are seen over elevated regions (e.g., the Alps), but also in some coastal areas (Norwegian coast). There are also areas with high return levels in the Mediterranean region as a consequence of humid air advection towards inland by cyclones coming from the sea.

Also the climate change signal was investigated in the RAIN project. The results show that the number of heavy precipitation events increases with increasing greenhouse gas concentrations [7], [8]. The highest increases were found in northern Scandinavia, western Ireland and western Scotland. The increase in the number of events was found both for the longer-lasting

[4] meteoalarm.eu

accumulative rain events and for the short-term high-intensity events, with the latter being more relevant.

Blöschl et al. (2017) [9] have recently studied the impact of changing climate on the timing of European floods by analyzing a large dataset of flood observations from the past five decades, 1960-2010. A clear shift in the timing of floods was found. Springtime flooding caused by melting snow has become earlier in northeastern Europe due to increasing temperatures, whereas earlier soil moisture maxima have led to earlier winter floods in Western Europe. Around the North Sea as well as in some areas of the Mediterranean coast, delayed winter storms associated with polar warming have led to late-winter floods.

2.2. Social Media in emergency context

Recently, the use of social media during emergencies and how it can be exploited to enhance situational awareness, has received much attention. In the work done by Olteanu et al. [10], the authors present the result of a crowdsourcing campaign aimed to describe what to expect from social media data across a variety of emergencies (natural disasters, terrorist attacks, explosions, etc.) in terms of volume, informative level, type and source. Twenty-six events have been considered, among which two Italian ones (one earthquake and one flood). A similar crowdsourcing approach has been used by the UK and Irish Met Offices [11]. Event detection from social media data was investigated in [12], where Sakaki et al. propose a system to automatically detect earthquakes in Japan using a probabilistic approach on the volume of Tweets, while Klein et al. [13] propose a Natural Language Processing (NLP) approach coupled with a clustering algorithm to tag Tweets as related to an emergency event or not. Similarly to ours, a lightweight volumetric approach is proposed in [14], where features are stored on a Cloud platform. Multivariate analysis is proposed in [15], but this method would pose a severe limits in using parallel computing to scale up the solution. An overview of semi-supervised methods for anomaly detection in time series can be found in [16]. Several works has been done concerning the classification of online data into information classes or topics. The closest work to ours on the emergency context is the one by Caragea et al. [17], which compares several approaches to classify text messages written during the Haiti earthquake and gathered by the Ushahidi platform[5] into different information classes. Another similar study is the one done by Asakura et al. [18], where a NLP techniques are used to understand whether a flood event has occurred taking into account also GPS information contained in the Tweets.

2.3. Novel Contributions

Our work is different because we propose a novel set of services that links meteorological forecasts with social media analysis. We propose a transdisciplinary methodology that exploits the availability of meteorological forecasts to (i) identify areas at risk and (ii) start a targeted monitoring through social media to acknowledge the occurrence of the forecasted weather events, (iii) detect associated natural hazards (floods in our study), and (iv) automatically filter the social media stream to retain only informative content. Here the concept of informativeness is defined as everything that can be useful to improve the situational awareness for both citizens and authorities about an emergency event. We envision two types of end-users for the proposed services. Firstly, hydro-met agencies (forecasters) who are interested to receive on-field observations as acknowledgments of model outputs. Secondly, first responders and local authorities who are interested to receive event detection alerts and relevant contextual information that can be exploited in order to understand the extent and criticality of an ongoing event when there is no personnel on the field.

3. The Case Study: flood in northern Italy

Twitter is the most studied social network in the emergency domain [5], probably due to the ease of sending and extracting information and to its open data policy. Twitter is categorized as a micro-blogging service, which is a form of communication that allows users to send brief text messages (formerly up to 140 characters, recently updated to 280), also known as Tweets, or media such as photographs or audio clips. By default, all user posts are public, and they can be automatically retrieved using Twitter's Application Program Interfaces (API), which can be freely used under the limitations specified in the terms of service [19]. As shown in Vieweg et al. [20], Twitter is also used to give situational information during emergency events: during

[5] https://www.ushahidi.com/

the Boston Bombing in 2013, it has been estimated that 27,800,000 Tweets were written about that event. Furthermore, the information provided by Twitter can very easily become viral (i.e., spread rapidly and on a vast scale across the Web) thanks to Retweets, which are generated when a user re-posts (forwards) a message from another user. For all the aforementioned characteristics, we select Twitter as the social media platform to investigate.

Among all natural hazards, flood is one the most devastating. The immediate consequences of floods are loss of human life, damage to property, and destruction of crops. Long-term consequences of floods include disruptions to supplies of clean water, psychological impacts, degradation of the electric power infrastructure, but also impacts on health care, education and environment. As has been analyzed by the U.S. F.E.M.A. (Federal Emergency Management Agency) [21], flood losses in the United States averaged $2.4 billion per year for the last decade, making flood the number one natural disaster in the United States.

Due to the aforementioned reasons, we select a weather-induced flood as our case study. However, the architecture of the proposed set of service is general and it can be easily extended to all hazards that depend on meteorological conditions, e.g., wildfires, landslides, avalanches.

We consider the flood in Northern Italy of November 2016, the details of which are fully available in the online official report [22] created by the Piedmont Region. The heavy rains fallen between November 22^{nd} and November 25^{th} in Piedmont (North-West Italian region) caused an significant flood, which mainly involved mountain areas and affected homes and infrastructures (roads and railways). On the 24^{th} and 25^{th}, the rainfall measured at stations near Turin reached over 50 mm per day. The event caused the evacuation of 1477 people in the affected areas, it left 350 people stranded and it caused, unfortunately, the death of a person. An alert was issued November 22^{nd} but the first reports were sent to the Civil Protection during the morning of November 23^{rd}. The first flood of the Tanaro river was reported on the 24^{th}, while during the day of November 25^{th} floods occurred also in the area of South Torino (Piedmont's chief town) causing the evacuation of 200 people. The flooding of the river Tanaro (the second longest river in Piedmont) happened again in the night between November 25^{th} and November 26^{th}, affecting the city of Alessandria and nearby municipalities. The relevance of this event is also confirmed from the Copernicus EMS activation (EMSR 192) that produced several delineation and grading maps [23].

4. A novel set of services to link early warning to emergency response

This section describes the user-centered set of services proposed within this paper, which aims to link the early warning to the emergency response phase coupling weather forecasts together with social media monitoring and analysis.

In our approach, social media analysis focuses on volume and textual features in order to allow a scalable and real-time analysis aimed at event detection and data extraction. Therefore, we leave out Social Network Analysis (SNA) on users communities because it would be computationally impractical, especially in case of large events that reach a world-wide news coverage.

The proposed set of services is composed of 4 different modules:

- Weather Forecast
- Social Media Monitoring
- Event Detection on Social Media Streams
- Informativeness Classification of Social Media Content

We assume that background social media monitoring jobs are always present in order to monitor the aggregated volume of content associated to a set of topics (in our work case extreme weather conditions and flood) and languages the end-user is interested in. The aggregated volumes are needed by the Event Detection module, as explained in Section 5.3. We also assume that end-users are allowed to define topics and languages, and that one monitoring job per topic-language is launched. The details about the topic definition and the social monitoring approach are given in Section 5.2. As shown in Figure 2, the process starts from the production of weather forecasts, which are used to identify areas that could be subject to extreme weather, i.e., areas at risk. If no area is found, the same check is performed again upon the generation of the subsequent forecast. Note that we assume that forecasts are operationally produced with a given periodicity by a meteorological agency. If at least one area is found, parallel instances of the monitoring and of the event detection algorithm are started, where each instance is related to a topic-language pair. The event detection algorithm outputs with a given temporal resolution a binary signal, i.e., *true* if an event is detected, and *false* if not. In the first case (*true*), an alert containing the data that triggered the algorithm is sent to the end user for verification, while the detection continues until the next forecast becomes available. If the end user confirms the presence of the event, the filtering task is started on the corresponding topic-language pair. Each tweet matching this pair received by the monitoring module is fed to a classifier (Section 5.4) that retains only informative content and shows them to the end user. If the event is not confirmed, the event detection algorithm continues after a freeze time. When the event is over, the end user notifies the system, which resumes from the event detection block after the freeze time. This approach requires that the system implements a user interface, e.g., a web application, in order to handle the data and signal exchanges with end users. Note that the reception and the subsequent validation of an event alert (e.g., flood) may be the responsibility of hydro-met agencies or of civil protection departments according to the regional/national division of competences. Even if the output of the filtering module mainly targets public authorities and first responders who have to manage the event, it can be relayed to any of the stakeholders involved.

5. Methodology

This section is focused on the detailed explanation of all methodologies we propose. We devote one subsection for each step of the process described in Section 4.

5.1. Extreme Weather Forecast

Accurate predictions of severe weather events are extremely important for the society, the economy, and the environment. Due to the fact that weather forecasts are inherently uncertain, it is required that information about forecast uncertainty be provided to all users, i.e., that weather forecasts are given in probabilistic terms. Weather forecast accuracy is limited by (i) the inaccurate description of the initial, observed state of the atmosphere and (ii) by the prerequisite to use approximations and simplifications in the actual weather forecast model equations. Furthermore, even the smallest uncertainties in the initial conditions of the forecast model have a tendency to grow rapidly with the lead time time because of the chaotic nature of the atmosphere. Therefore, rather than integrating a single forecast from a supposedly best guess of the initial state, a better approach consists in starting the forecast from a number of slightly different initial conditions, and then deriving as many outcomes from these initial conditions (Palmer, 2000 [24]). This approach is called ensemble forecasting, and it outputs forecasts as probability distributions, from which local probabilities can be computed for different weather events by using thresholds. Similarly to what is operationally done by the most advanced weather centres, e.g., the ECMWF (European Centre for Medium-Range Weather Forecasts), we propose to run operationally ensemble forecasts twice a day.

5.2. Social Media Monitoring

Today we are using an unprecedented wealth of social media platforms to share information about everything that is happening around us. In the emergency domain such information can become a powerful resource for assessing in near real-time the evolution of an hazardous event, its impact and how it is perceived by the affected population. Hence, the goal of the social media monitoring module consists in retrieving content related to selected hazards in order to extract contextual information that could be useful to citizens, forecasters, first responders, and decision makers.

The social media platform monitored in our case study is Twitter, because it is a news-oriented social network and it has been used in many previous studies in the emergency domain (see: [20, 12, 13, 10]) that exploit and analyze Twitter content. Furthermore, Twitter data are openly accessible through public APIs.

The monitoring process is triggered by the detection of an extreme weather event (possibly encoded in a hydro-meteo bulletin) that defines the geographical regions at greater risk and the hazards to be monitored. Note that the monitoring is activated only on the language of the regions identified by the forecast and on predefined set of keywords, one for each of the considered hazards.

To retrieve social media content, the Social Media Monitoring (SMM) modules relies on the Streaming API exposed by Twitter [25]: such APIs are designed to follow specific topics (or users) enabling low latency access to Twitter's global stream of data by pushing messages, thus avoiding the overheads associated with polling an API endpoint. However, these public, cost-free, Streaming APIs are characterized by an overall volume limitation of 1% (randomly subsampled) of the total stream (see [19]), i.e., whenever the volume of a filtered stream is greater than 1% of the total stream.

In order to avoid this subsampling and maximize the volume of retrieved relevant content, it is important to limit the off-topic content by configuring the access to the global Twitter stream with one of the different filtering parameters exposed by the Streaming APIs. The main options that the Streaming API allows to filter the content are:

- **language**: the language of the content;
- **locations**: one or more geographical regions, identified by their bounding box (if set, only geolocalized tweets are retrieved);
- **follow**: a list of authors ID;
- **track**: a set of terms (words or hashtags) that should be present in the content. A track phrase includes one or more terms (separated by spaces) and a match is returned if at least one of the phrases is present in the Tweet, which will then be delivered to the stream.

Among these filtering parameters, the SMM exploits the **language** and the **track** phrases. The **follow** parameter is not pertinent in our use case, since the module aims at retrieving all the content related to a topic regardless of the author. Instead, the **locations** parameter would be too restrictive because it retains only the geolocalized Tweets, which are less than 2% of the posts ([26]). The monitoring module is configured with a set of track phrases: one for each of the supported hazard types and languages.

Track phrases are textual queries, expressed in a simple syntax: no exact matches or exclusions are possible. The content in each monitored stream is then processed through a Language Analysis pipeline (involving lemmatization, key phrases detection, named entity recognition, classification and sentiment analysis) that enriches them with additional linguistic and semantic metadata; more details on the used pipeline can be found in [27]. After this pipeline, a second and more refined one is applied in order to filter out unrelated content (e.g., texts such as “landslide victory”, “flood of votes”, etc.). These classification rules are based on language and semantic features (e.g., lemmatization, proximity expressions, exclusions) and are manually composed by mother tongue domain experts. One set of rule for each topic (hazard) is required.

The final output is stored in a database to be exploited by the other services, i.e., the Event Detection and the Informativeness Classification (Filtering) modules.

Note that, regardless of the monitoring processes activated by weather forecasts, simple monitoring processes (one for each of the considered event) is always present in order to compute the volume of social media content grouped by language and event type in a given time window. This aggregated data is stored in the database and exploited by the Event Detection module (see Section 5.3).

5.3. Event Detection

In this subsection we describe the proposed algorithm for event detection designed to detect emergencies, or anomalous phenomena. The Event Detection Module (EDM) analyzes streams of data that are generated by the SMM component. These streams are differentiated by language and topic (event type / phenomenon), as described in Section 5.2. This brings three main advantages to the event detection procedure:

1. it removes unnecessary noise that might hinder the detection of a specific phenomenon;
2. it provides a basic description of the event which is unfolding. An extreme weather forecast can potentially be related to several events (e.g., storms, floods, landslides);
3. in some languages (such as Italian) most Tweets will be originated from the interested country. This helps in filtering relevant content, as the chance that posts apply to a local emergency are higher. By comparison, it is more difficult to understand if an English Tweet about floods relates to an Italian emergency.

One of the main requirements of the EDM is to properly handle heterogeneous data with respect to:

- content: not only different events types, but also several events of the same type, as they might have very different behaviors ([10]);
- type of emergency: some emergencies can be forecasted (e.g., flood) while others cannot (e.g, earthquakes), which translates into content related to the monitored hazards being available at different time scales;
- volume: because the extent of emergency events in terms of affected people and geographical area can be very different, the volume of the generated social media content varies too. Furthermore, it also depends on the social media adoption (active users) in the affected area.

We propose a volume-based EDM that operates on the series of tweets, aggregating them in predetermined time-frames. As mentioned in Section 5.1 the system does not keep a copy of each Tweet, unless it has been collected in relation to a validated event, while the aggregated volumes per time-frame are stored indefinitely for further analyses and tuning of the EDM.

Our EDM builds upon the generalized Extreme Studentized Deviate test (ESD) ([28], [29]). We consider a discrete and integer time scale, where each time slot has the same size S. Given a language l and an event type e, let $\overline{X}_{l,e}$ be the time series (stream) of volumes related to l and e. Once a new element of the time series $X_{l,e}(t)$ is added at time t, the series is tested for outliers within a sliding window $w = (t - W, t]$, where W is the number of time slots. If $X_{l,e}(t)$ is considered an outlier, the alert is triggered. Therefore, the system works in near real-time, with a periodicity of S and, because it works on univariate time series, is also asynchronous on different streams. The basic ESD test is improved as in [30], where a volume-based method on heterogeneous Twitter production data has been developed and tested. This procedure takes into account seasonality in Twitter activity by using time series decomposition, which allows to detect local anomalies (inside seasonal patterns) on top of global anomalies (which are easier to identify). The most important patterns in the considered scenario are the day/night one and the weekend one. This technique avoids, for example, to under-report night events.

The algorithm also employs the median instead of mean in the original ESD test, making it statistically more robust. This allows to properly account for low-volume data, for example in events happening in sparsely populated regions, where the Twitter community is smaller. Note that if the activity is near zero even in the emergency phase, the system can not be effectively used to provide early warning signals on smaller time-frames. However, relevant events are usually reported well beyond the original impact area, helping the detection module to trigger despite the low affected population.

The anomaly detection procedure follows a two-step schema. If an event is detected, a summary of the content that triggered the algorithm is generated and forwarded together with the alert to help first responders in assessing and validating the detection. The summary considers all tweets at time t plus aggregated measures (see Section 6.3 for details). As explained in section 4, after one alert is sent a freeze time F in terms of number of slots is set before running the next detection. Also, as long as the event is in progress no more alerts are pushed (for that stream). When the event is declared to be over, relevant historical data are saved and kept for future uses. Additionally, if the system incorrectly signals an event, the EDM is frozen for F in order not to provide first responders a series of false positives. In such case, no detailed data are saved.

5.4. Informativeness Classification

The objective of this component consists in classifying informativeness from tweet texts, thus classify tweets in "informative" and "not informative" classes. What is considered as informative depends on the user of the information, as such is considered as an arbitrary concept. In this study we defined informativeness as in [10], thus considering as informative all contributions that are relevant to the crisis situation and at the same time help to improve its understanding. Hence, Tweets in which the crisis situation is mentioned but do not contain information that is helpful to understand it are not considered informative. In order to capture informativeness we consider Natural Language Processing (NLP) techniques based on vector representation of the Tweet text. In particular we focused on the `fasttext` [6] tool [31, 32, 33] developed in the Facebook AI Research group. In this approach n-grams are learned instead of words and a word is seen as a sum of n-grams. This method can be seen as an extension of the continuous skipgram model [34] because it takes into account sub-word informations. We choose this model in view of its performances in the analysis presented in `fasttext`, where it has shown good accuracy (comparable with other methods) and at the same time a faster learning process [32].

The ability to discriminate between "informative" vs "not informative" tweets is a document binary classification problem. The standard metrics used in such cases are **precision**, **recall** and **f-1** score. In this case, we focus on the performance of the algorithm for retrieving "informative" data. All the metrics refer to this specific class of tweets. In this sense, **precision** is the ability of the classifier not to label as "informative" a sample that is "not informative", **recall** is the ability of the classifier to find all the "informative" samples and **f-1** represent a sort of harmonic average of the two.

[6] https://github.com/facebookresearch/fastText

6. Implementation

The following subsections provide details on the implementation of each module.

6.1. Weather Forecast

Ensemble-based early warning products have been developed to forecast severe weather events by utilizing both the ECMWF-ENS (European Centre for Medium-Range Weather Forecasts Ensemble prediction system, 51 members) and GLAMEPS (Grand Limited Area Ensemble Prediction System, 52 members) models. These products estimate the occurrence probabilities of heavy rainfall, strong winds, and extreme high/low temperatures, and it is routinely produced for the whole European area. The forecasted occurrence probability of the different severe weather events is computed according to pre-defined thresholds. However, as was highlighted in Section 5.1, for weather forecasting and warning services a fixed threshold is suitable and it is typically used in operational forecasting globally. Here, when studying the skill of the forecasts in the Piedmont case, we have used 50 mm as the threshold for the 24h rainfall, which is at the lower boundary of the range of values, 50-100 mm/24h, attained from the stakeholder interviews carried out in the RAIN project [6].

Since ensemble forecasts are typically under-dispersive and/or biased they should be calibrated by utilizing statistical methods. If forecast system is under dispersive, the range of possible ensemble solutions is too small compared to what frequently happens. Bias can be either positive or negative, and it means that the predicted ensemble mean is systematically either larger or smaller than observation on average. Most of the recently used statistical methods share a general approach of correcting the current forecast by using past forecast errors, as has been done for deterministic forecasts in the so-called Model Output Statistics (MOS) procedure introduced originally by Glahn and Lowry (1972) [35]. This process makes use of information from prior forecasts and observations to produce probabilistic forecasts or to improve their reliability. The method providing the best outcome is dependent on the weather variable being forecasted. Statistical calibration is found to be useful at a variety of time scales including short forecast lead times, and even lead times of up to two weeks.

6.2. Social Media Monitoring

The monitoring module was implemented in Java and Scala as a job within the Spark Streaming architecture [36], which is an open source infrastructure designed to deal with real-time data analysis, transformations and operations. CELI proprietary resources were used in the Language Analysis pipeline [27]. Storage was performed on PostgreSQL, which is a well known open source database [37].

In the proposed case study the language of the monitored content is Italian and the set of keywords used as track phrases for filtering the social stream consist of:

- **Flood specific keywords**: alluvione, alluvioni, esondazione, esondazioni, esondato, esondata, esondare, allagato, allagata, allagamento, allagamenti, "cedimento argini", "cedimento argini", "ceduto argine", "cede argine", "ceduto argini", "cedono argini", angelidelfango, inondazione, inondazioni
- **Weather related keywords:** maltempo, allertameteo, meteo, pioggia, piogge, piove, piovere, piovuto, piover, nubifragio, nubifragi, "bomba d'acqua", "bombe d'acqua", bombadacqua, bombedacqua, allarmemaltempo
- **Other hazard related keywords:** slavina, slavine, smottamento, smottamenti, idrogeologico, idrogeologici, frana, frane, franare, franato, franata, franate, franati

These keywords generates what we define as the *Monitoring* stream.
The fine grained classification rule operating on the use case data, defining the *Event Detection* stream, is:

- maltempo [alluvione] [esondare] [allagare] [inondare] "[cedere] [argine]" ~4 angelidelfango -([voto] [politica] [elezioni])

Terms enclosed between [ ] match on the lemmatized form of the textual content (i.e., [allagare] is a verb and it will match on any form and tense of that verb). Terms preceded by a minus sign represent term that should not be present in the retrieved content. Expressions followed by a tilde and a number N are proximity expressions that identify documents containing all the terms in the expression, each one within a maximum distance of N terms between the others.

Neither the track phrases nor the classification rules contain any reference to a specific geographical entity. This allows the component to work independently from the location of the hazard, so it can be used to detect new events without having a mandatory a named entity recognition/disambiguation component.

6.3. Event Detection

We implemented the EDM in the R language, using the AnomalyDetection[7] for the the ESD, and SparkR[8] for accessing an Apache Spark cluster, which is used to computed the volumetric measure. Tweet streams are stored in the PostgreSQL and aggregated according to the language l, and event type e.
For each l and e, we store a 1 hour time series $\overline{X}_{l,e}$ and a 15-minute time series $\overline{Y}_{l,e}$. $\overline{X}_{l,e}$ is tested regularly (every hour), checking if the new entry $X_{l,e}(t)$ is an outlier. If an outlier is detected at t_0 using $\overline{X}_{l,e}$, we start testing $\overline{Y}_{l,e}$ every 15 minutes and only after the outlier is confirmed also on time series at time $(t \geq t_0)$ we issue the event detection alert. This is done to reduce the computational cost.

Anomaly detection is implemented using the Seasonal Hybrid ESD Test ([30]), which depends on four parameters p, w, th, α.

- p is the piecewise median time window; we set $p = 2$ weeks, the minimum value allowing to take into account weekly periodicities, such as the weekend effect;
- W is the sliding window size (mentioned in section 5.1) that we set to 2 weeks, which is the minimum value to support $p = 2$ while keeping the procedure lightweight;
- th: is the threshold for setting the percentile of the daily max values used to trigger the anomaly detection. We set $th_x = 0.95$ (95th percentile) and $th_y = 0.99$ (99th percentile) for $\overline{X}_{l,e}$ and $\overline{Y}_{l,e}$, respectively.
- α: minimum level of statistical significance for anomalies; we set $\alpha = 0.01$. In our case studies, it is far less determinant than th. Statistical significance can be used by further components of the system (see below).

[7]https://github.com/twitter/AnomalyDetection
[8]https://spark.apache.org/docs/latest/sparkr.html

We have empirically set such parameters in order to achieve the best accuracy on a wide data set comprising 280k tweets collected during 3 Italian emergencies (snow, earthquake, landslides) from Oct. 2016 to Jan. 2017. p and W can be safely increased depending on the computational power at disposal, as more data must be collected and analyzed with greater W and p. We did not register significant variations of the event detection accuracy with W. Conversely, th_x and th_y control the trade-off between Precision and Recall of the detection. In emergency management it is desirable to favor Recall compared to Precision, as (potential) disasters should never be missed in the detection phase. Hence, we choose a lower th_x in order to keep the EDM sensitive enough, and an higher th_y to precisely pinpoint the emergency event. If an anomaly is detected, the R module returns the binary signal *true*/*false*, the confidence level, and relevant metadata, i.e., the identifiers of the stream (including language l and event type e) and the time-stamp corresponding to the end of the slot that triggered the detection.

6.4. Informativeness Classification Implementation

To test the performance of `fasttext` in classifying informativeness of Tweets collected during emergencies, we look at the CrisisLexT26[9] database [10], a collection of tweets collected during 26 different crisis situations, which took place between years 2012 and 2013 at different locations of the world. All collected Tweets have been manually labeled by local citizens in different classes with respect to the event under study, i.e., “related and informative”, “related and not informative”, “not related”, and “not applicable”. We aim to implement a binary classifier that detects the class “related and informative”. Hence, this class is our positive class and everything else is discarded, in other words, everything else is our negative class. This dataset presents two main difficulties with respect to other Twitter datasets used for text analysis: the corpus of labeled data is small (less than thousand tweets for each event, see Table 1) and the dataset is in different languages. The balance of the analyzed Tweets (see Table 1) is in several cases greater than 0.5, the optimum balance for training of the classifier. However, we do not see any evidence that this imbalance influences the final performance, except in two cases (NY train crash and Philippines flood), thus we consider the data to be a valid training set.

[9] http://crisislex.org/data-collections.html

Event Name	**Num tweets**	**Balance**
Bohol earthquake	671	0.50
Boston bombings	658	0.47
Brazil nightclub fire	589	0.54
Colorado floods	804	0.81
Glasgow helicopter crash	688	0.56
LA airport shootings	738	0.68
Lac Megantic train crash	618	0.58
Manila floods	675	0.64
NY train crash	658	0.89
Queensland floods	807	0.73
Colorado wildfires	957	0.60
Russia meteor	881	0.47
Sardinia floods	744	0.61
Savar building collapse	456	0.55
Singapore haze	543	0.45
Spain train crash	656	0.81
Typhoon Yolanda	751	0.72
West Texas explosion	683	0.52
Costa Rica earthquake	1051	0.50
Guatemala earthquake	743	0.73
Italy earthquakes	737	0.66
Philipinnes floods	551	0.84
Typhoon Pablo	742	0.71
Venezuela refinery	766	0.57
Alberta floods	786	0.72
Australia bushfire	885	0.62

Table 1: CrisisLexT26 dataset analyzed in terms of number of Tweets and balance, where balance means informative Tweets versus all.

7. Results

In this section we present the results achieved assessing the proposed set of service with the selected case study.

7.1. Weather Forecast

Quality developments of numerical weather prediction models are good indicators of forecast usefulness and applicability in different time scales. The predictability of the ECMWF model based precipitation was of the order of 2 days in the mid-1990s and had increased up to 3.5 days by 2010 (Nurmi, et al. 2013 [38])). The trend in predictability improvement has been fairly linear during past decades with an increase of about one day per decade. Therefore, the predictability of precipitation is expected to improve also in the foreseeable future at a relatively constant rate and is today around 4 days.

The heavy precipitation event of our Piedmont case study (see Section 3) was well forecasted by the ECMWF ensemble model and is in good agreement with the above discussion. The probabilistic forecasts of accumulated precipitation exceeding 50 mm during a 24-hour period (from 23rd of November 18 UTC to 24th of November 18 UTC) in the Piedmont area (and southern France) can be seen in Figure 3. The figure shows four different forecast cycles made 234, 162, 90 and 30 hours ahead of the forecast valid time of 24th of November (18 UTC) to highlight the forecast evolution with respect to the forecast lead time. In this particular case, the forecasted heavy rainfall probabilities were higher than 70% already as early as six days (162 hours) before the event, thus providing remarkably early warning guidance against an upcoming event. This is about two days earlier compared to the average precipitation forecast skill (approx. 4 days) explained above. Three days (90 hours) before the event, the forecasted probability of heavy rain was very high (between 90 and 100%) over large areas in the Piedmont region. When heavy rainfall is predicted to occur over densely populated area, like in this case, it is common practice to initiate actions when heavy precipitation probability is over 50%.

7.2. Social Media Monitoring

With the configuration described in Section 6.2 the monitoring module collected 92,760 elements in the considered time range (from 19th to 26th of November, 2016). A dataset containing the raw JSON of these Tweets (a textual file containing in each line the JSON serialization of a single Tweet) has been released as a public resource[10] with a Creative Commons license. The published dataset consists only of the raw JSON (as it is provided by Twitter) and not the enrichments, i.e., metadata and classification computed.

The published dataset contains:

- 52,349 original Tweets (not considering Retweets)
- 1,234 Tweets containing an exact localization (a point, based on the device GPS)
- 2,150 Tweets containing an approximate localization (a bounding box, based on the device network connection)
- 14,995 Tweets containing a photo, 7,181 of which unique (not considering re-posts of the same photo)

The distribution of the Tweets volume in the considered time range is reported in Figure 4. The tweets containing an exact localization are fully plotted on a map of Italy, reported in Figure 5. An additional map (Figure 6) shows the regions affected by the flood. In Table 2 we show the top 10 most province by number of Tweets per population, which we compute according to the province area (NUTS 3). We note that such frequency of Tweets in an area are not sufficient information to determine the origin of the event.

7.2.1. Dataset Content Overview

In this Subsection we show a qualitative representation of the content in the dataset and how it evolved during the event, by visualizing for each day the most frequent key phrases. Key phrases are extracted by the language analysis pipeline identifying specific patterns of terms with desired linguistic features (i.e. a noun followed by a preposition and another name or an adjective followed by a noun). Word Clouds are then computed by selecting the most frequent key phrases within a given time period.

Figure 7 represents a Word Cloud of the most frequent key phrases extracted on the 23rd of November, while 8, 9 and 10 on the following 3 days.

It can be observed that the main topics emerging from the Word Cloud computed on the 23rd of November are related to alerts (“allerta meteo”,

[10] https://www.zenodo.org/record/854385/files/PiemontFlood2016Dataset.zip

#	Region	Province	Tweets (T)	Population (P)	T/P ‰
1	Umbria	Terni	103	229 071	0.4496
2	Liguria	La Spezia	60	221 003	0.2715
3	Calabria	Crotone	41	174 712	0.2347
4	Lombardia	Milano	705	3 208 509	0.2197
5	Liguria	Savona	56	280 707	0.1995
6	Apulia	Taranto	87	586 061	0.1484
7	Piemonte	Cuneo	71	590 421	0.1203
8	Piemonte	Torino	255	2 282 197	0.1117
9	Lazio	Roma	468	4 340 474	0.1078
10	Piemonte	Biella	18	179 685	0.1002

Table 2: Top 10 most province by number of Tweets per population, computed according to the Italian NUTS 3 level.

"allerta arancione", "allerta rossa") and heavy rains ("forti piogge", "forti precipitazioni", "pioggia in aumento", "forte maltempo", "temporali e schiarite"). On the 24th the focus is both on maximum alert levels ("allerta massima", "allarme maltempo", "allerta arancione") and on flood warnings and locations/rivers ("incubo alluvione", "Piemonte e Liguria", "fiume Tanaro", "Tanaro nel cuneese", "fiume Po", "Po a torino", "Tanaro in piena"). On the 25th, instead, the alerts topic is almost disappeared while floods and rivers topics are still present ("piena del Po", "piena a Torino", "esondazioni Piemonte") as well as other themes related to the emergency caused by the flood ("Renzi a Torino", "scuole chiuse", "videoclip ufficiale"). Finally ,on the 26th the topics emerging from the Word Cloud include other events/locations ("maltempo in Sicilia", "Po in Lombardia") besides the flood in Piedmont and Liguria ("Tregua in Piemonte", "frane in Liguria").

From this overview we can conclude that key phrases can be a useful instrument in order to assess the presence of given meteorological events, like floods, as well as the affected locations and rivers.

7.3. Event Detection

As described in Section 6.2, in our case study we considered : a more generic Monitoring, which combines extreme weather and floods, and a flood-specific one that was used for the Event Detection.
A comparison between these two streams is reported in Figure 11, where it

can clearly be seen the difference in volumes, especially in the days prior to the emergency.

We plot $\overline{X}_{l,e}$ and $\overline{Y}_{l,e}$ together with the positive signals (detected anomalies) with l='it' and e='flood' in the period surrounding the event in Figures 12 and 13, respectively. Since no end-user was actually involved in the evaluation, we show all detections. The hourly test for anomalies in $\overline{X}_{\mathrm{it,flood}}$ is first passed with 455 Tweets on the 24$^{\mathrm{th}}$ of November at 10 UTC (11 CET). These Tweets were posted between 9 and 10 UTC. The subsequent test on $\overline{Y}_{\mathrm{it,flood}}$ is first passed at 10.30 UTC (11.30 CET).

This result corresponds to the maximum alert level, which was reached between morning and afternoon of the 24th of November, as mentioned in the official report by the Civil Protection. The exact time of the first flood is not mentioned in any official records at local/regional level, while the Event time reported by Copernicus EMS, is November 24th, 17.00 UTC (18.00 CET). Most frequent hashtags used in the flood stream in the hour of the alert were: #tanaro (131 tweets), #maltempo (115), #piemonte (54), #liguria (27), #allertameteopie (23). #piemonte, #allertameteopie, #liguria relates to the regions interested by the event (Piedmont and Liguria), while Tanaro is the main river whose waters caused the hazard. Most frequent named entities (locations) detected in the same hour were: Piedmont (212 tweets), Liguria (135), Tanaro River (125), Province of Turin (33), Garessio (19). Garessio is the town originally affected by the Tanaro river flood. We can see that a ranking of the tops hashtags and named entities is useful to spot the location of the flood.

According to official statements [11], first responders received damage reports from on-field agents during the course of the 24$^{\mathrm{th}}$ of November. Timing of the first social media alert appears to be in line with these reports, as Tweets concerning the Tanaro flooding were generated immediately after witnessing the hazard, prompting the warning system. In this case, both social media and on-field agents reports provided a quick alert. We do not claim that a social media early warning system, such as the one we propose, is necessarily more reactive than on-field personnel. However, the social media detection is extremely useful in case those agents are missing.

[11] http://www.regione.piemonte.it/cgi-bin/montagna/pubblicazioni/frontoffice/richiesta.cgi?id_settore=10&id_pub=1394&area=10&argomento=111

7.4. *Informativeness Classification*

7.4.1. *CrisisLexT26 dataset Result*

In Table 3 we show the performance of `fasttext` in classifying informativeness of the 26 crisis events. In order to maximize the amount of data in the learning phase, the analysis was performed for each single event using all Tweets from the other 25 events for training. Thus, we have used the so called "leave-one-out" approach [39] in order to test the performance of the proposed method. Even if this reflects in learning simultaneously by using diverse languages, the obtained accuracy is greater than considering only Tweets in a single language, which corresponds, for each event, to the mother tongue of the country in which the event occurred.

We obtain an average f-1 score of 78% on natural hazards, which shows the efficacy of the method. From the point of view of languages we can see that the Italian and the Russian language seem to be particularly challenging for the considered task. In the following, we will see how to improve these results for the Italian case study.

7.4.2. *Flood in Piedmont*

After the successful test with the CrisisLexT26 database we move forward to analyze the tweets collected before and during our case study, i.e., the flood in Piedmont. In particular, we choose around 1200 Tweets and manually annotate them according to informativeness. The considered Tweets were generated starting from the 19th until the 26th of November, 2016.

In order to test the `fasttext` algorithm for this selection of Tweets we need to identify an appropriate corpus of tweets for the training phase. Having at our disposal the CrisisLexT26 database, we decided to test how performance changes using different subsets of this database. In particular, we train the algorithm in three different ways:

- Italian: using only tweets connected with Italian emergencies;
- Nat. hazards: using only tweets from natural hazards;
- All: using all the tweets from CrisisLexT26.

With the aim of early detection and monitoring of emergency events, we choose to not consider Tweets as a unique body but instead we group the Tweets generated in time intervals of two hours (the minimum time span for

	Precision	**Recall**	**f-1**
Bohol earthquake	**0.90**	**0.81**	**0.85**
Boston bombings	0.85	0.51	0.64
Brazil nightclub fire	0.88	0.47	0.61
Colorado floods	**0.91**	**0.81**	**0.86**
Glasgow helicopter crash	0.81	0.72	0.76
LA airport shootings	0.87	0.78	0.82
Lac Megantic train crash	0.76	0.68	0.72
Manila floods	**0.87**	**0.75**	**0.81**
NY train crash	0.96	0.90	0.93
Queensland floods	**0.87**	**0.76**	**0.81**
Colorado wildfires	**0.79**	**0.77**	**0.78**
Russia meteor	0.72	0.44	0.54
Sardinia floods	**0.90**	**0.41**	**0.56**
Savar building collapse	0.68	0.69	0.68
Singapore haze	0.68	0.63	0.65
Spain train crash	0.93	0.74	0.83
Typhoon Yolanda	0.87	0.75	0.80
West Texas explosion	0.84	0.72	0.77
Costa Rica earthquake	**0.75**	**0.87**	**0.81**
Guatemala earthquake	**0.92**	**0.83**	**0.87**
Italy earthquakes	**0.91**	**0.44**	**0.59**
Philipinnes floods	**0.91**	**0.88**	**0.89**
Typhoon Pablo	0.91	0.82	0.86
Venezuela refinery	0.80	0.47	0.59
Alberta floods	**0.88**	**0.69**	**0.77**
Australia bushfire	**0.83**	**0.77**	**0.80**
Average	0.85	0.70	0.75
Average nat. haz.	**0.87**	**0.73**	**0.78**

Table 3: Performance of the fasttext algorithm on the CrisisLexT26 dataset. Natural hazard events are in bold.

a sufficient data flow) and test the algorithm on each subset. Within this choice the average balance in the two classes for the intervals is 0.53%.

The results are presented in Figure 14 and summarized in Table 4. The figure immediately highlights the difference between Italian emergencies only and the other two cases. In the "only-Italian" case we have a very high recall due to a poor precision performance: what happens is that the algorithm considers almost everything as informative. The balance between the two metrics improves when the corpus includes Tweets in all languages instead of a single one (Italian, in this case). This can be understood in terms of enlargement of the vocabulary. It is highly probable that, in Tweets regarding an Italian event, terms in foreign languages (e.g., English) appear. Thus, the performance of the classifier is improved by including Tweets in other languages in the training set. Moreover, focusing on Tweets connected with natural hazards helps. From the point of view of the ability of early detection we do not see any significant transition in the effectiveness of the algorithm capturing informativeness before or after the event.

We also specifically checked the percentage of geolocalized tweets that are informative in the Piedmont dataset because a geolocalized *and* informative Tweet could be especially useful. An additional manual annotation was performed on the 1,234 geolocalized Tweets to assess their relatedness and informativeness to the crisis. However, only 26 of them were related to the crisis, and only 18 of this subset is informative. We discovered that most geolocalized Tweets were generated by weather stations and contained weather reports, hence not referring to a specific event. In this case, geolocalized Tweets were mostly useless for end users, who already have data from weather stations at their disposal.

Summarizing the average results, reported in Table 4, we can say that the overall effectiveness of the method is close to 70% and that the performance is strongly connected with the training dataset. In this sense, we can expect a performance improvement by labeling additional data of future natural hazard emergencies.

8. Conclusions and future works

8.1. Conclusions

We have shown how social media data can be used together with probabilistic weather forecasts to automatically detect an ongoing event and to

	Precision	Recall	f-1
Italian	0.60	0.87	0.68
Relevant	0.72	0.72	0.69
All	0.72	0.67	0.66

Table 4: Different performances of the `fasttext` algorithm on Tweets connected with the flood in Piedmont, using different training sets.

extract useful information. We used key phrase extraction as qualitative confirmation tool for weather forecasters, while the event detection plus the informativeness classification could be effectively used by emergency responders. Our results show that machine learning methods trained on data generated within past emergency events can generalize well on new data, which confirms the validity of our approach given the ever-changing nature of considered disasters.

8.2. Approach Limits and Future Work

The presented version of the monitoring module collects social media data by purely leveraging on their textual content, without any consideration on authors accounts and if they should be trusted or not. This approach is potentially open to undesired content from fake accounts (i.e., bots and trolls). Other undesired data might be retrieved as well by the monitoring module when a track keyword is used outside of the emergency context. Ideally, the Language Analysis pipeline should filter them out, but since the filtering process is prone to errors, a certain number of undesired contents is bound to remain in the collected dataset.

We decided to leave this issue out because in normal conditions such out-of-context data are continuously distributed over time and do not concentrate in a short period. Hence, they do not impact the Event Detection Module. However, investigating how to mitigate the effect of undesired content and/or fake accounts might be an interesting point for future research. Filtering layers could also be added or existing ones could employ more selective disambiguation rules. In the current system, we chose to favor Recall in detecting critical situations over Precision, as end users have to provide the final form of validation. Our work is based on the assumption that the Twitter data is freely available. However, should it be no longer the case in the future, a cost-benefit analysis should be performed in order to assess the sustainability of the proposed solution.

As future works we also intend to extend our analysis on more hazards and different languages, and exploit the use of image analysis. Deep learning techniques used for classifying images extracted from social media posts could contribute to both the Event detection (by adding up to existing text-based volumes) and the filtering (in case additional information is provided in the form of a photo). Available services such as Google Vision or Microsoft Cognitive Services could be used by a separate module that classifies the images, and a new study would be required to assess the performance improvements, if any. This approach could leverage widely used image-center social media like Instagram, as a data source for novel emergency management services.

Acknowledgements

This work was partially founded by the European Union through the I-REACT project (H2020-DRS-1-2015), grant agreement n.700256.

[1] NASA, SVS NASA Climate Change, `http://svs.gsfc.nasa.gov/vis/a000000/a004100/a004135/index.html`, 2012. [Online; accessed 19-July-2017].

[2] R. Pachauri, L. Meyer, Climate Change 2014: Synthesis Report, Technical Report, IPCC, 2014.

[3] T. KE, Changes in precipitation with climate change, Climate Research 47 (2011) 123 – 138.

[4] E. Commission, Copernicus emergency management service, `http://emergency.copernicus.eu/mapping`, 2017. [Online; accessed 19-July-2017].

[5] T. Simon, A. Goldberg, B. Adini, Socializing in emergenciesa review of the use of social media in emergency situations, International Journal of Information Management 35 (2015) 609 – 619.

[6] P. Groenemeijer, N. Becker, et al., Past cases of extreme weather impact on critical infrastructure in europe, `http://rain-project.eu/wp-content/uploads/2015/11/D2.2-Past-Cases-final.compressed.pdf`, 2015. [Online; accessed 30-August-2017].

[7] K. M. Nissen, U. Ulbrich, Will climate change increase the risk of infrastructure failures in europe due to heavy precipitation?, Nat. Hazards Earth Syst. Sci. Discuss (2016).

[8] P. Groenemeijer, A. Vajda, et al., Present and future probability of meteorological and hydrological hazards in europe, http://rain-project.eu/wp-content/uploads/2016/09/D2.5_REPORT_final.pdf, 2016. [Online; accessed 30-August-2017].

[9] G. Blöschl, J. Hall, et al., Changing climate shifts timing of european floods, Science 357 (2017) 588–590.

[10] A. Olteanu, S. Vieweg, 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 and Social Computing, CSCW '15, ACM, 2015, pp. 994–1009.

[11] U. M. Office, Name Our Storms 2016, https://www.metoffice.gov.uk/news/releases/2016/nameourstorms2016, 2016. [Online; accessed 01-January-2018].

[12] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th international conference on World wide web, WWW '10, ACM, 2010, pp. 851–860.

[13] B. e. a. Klein, Emergency Event Detection in Twitter Streams Based on Natural Language Processing, Springer International Publishing, pp. 239–246.

[14] C. Wang, K. Viswanathan, L. Choudur, V. Talwar, W. Satterfield, K. Schwan, Statistical techniques for online anomaly detection in data centers, in: Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on, IEEE, pp. 385–392.

[15] J. Kline, S. Nam, P. Barford, D. Plonka, A. Ron, Traffic anomaly detection at fine time scales with bayes nets, in: Internet Monitoring and Protection, 2008. ICIMP'08. The Third International Conference on, IEEE, pp. 37–46.

[16] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM computing surveys (CSUR) 41 (2009) 15.

[17] C. Caragea, N. McNeese, A. Jaiswal, G. Traylor, H.-W. Kim, P. Mitra, D. Wu, A. H. Tapia, L. Giles, B. J. Jansen, J. Yen, Classifying text messages for the haiti earthquake (2011).

[18] Y. Asakura, H. Masatsugu, K. Mamoru, Disaster analysis using user-generated weather report, in: Proceedings of the 2nd Workshop on Noisy User-generated Text, ACL, 2016, pp. 24–32.

[19] Twitter, Twitter API limits, `https://dev.twitter.com/rest/public/rate-limiting`, 2017. [Online; accessed 19-July-2017].

[20] S. Vieweg, A. L. Hughes, K. Starbird, L. Palen, Microblogging during two natural hazards events: What twitter may contribute to situational awareness, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, ACM, 2010, pp. 1079–1088.

[21] FEMA, Mapping The Risk: Flood Map Modernization, `https://www.fema.gov/pdf/about/regions/regionv/faq_east_stlouis.pdf`, 2017. [Online; accessed 10-August-2017].

[22] P. C. R. Piemonte, Flood Report, `http://www.regione.piemonte.it/alluvione2016/dwd/rapporto_evento_nove2016.pdf`, 2016. [Online; accessed 01-January-2018].

[23] C. EMS, Emsr192, `http://emergency.copernicus.eu/mapping/list-of-components/EMSR192`, 2017. [Online; accessed 19-July-2017].

[24] T. N. Palmer, Predicting uncertainty in forecasts of weather and climate, Rep. Prog. Phys. 63 (2000) 71–116.

[25] Twitter, Twitter Streaming API, `https://dev.twitter.com/streaming/overview`, 2017. [Online; accessed 20-August-2017].

[26] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, E. Shook, Mapping the global twitter heartbeat: The geography of twitter, First Monday 18 (2013).

[27] F. Tarasconi, V. Di Tomaso, Geometric and statistical analysis of emotions and topics in corpora, IJCoL - Italian Journal of Computational Linguistics 1 (2015).

[28] B. Rosner, On the detection of many outliers, Technometrics 17 (1975) 221–227.

[29] B. Rosner, Percentage points for a generalized esd many-outlier procedure, Technometrics 25 (1983) 165–172.

[30] O. Vallis, J. Hochenbaum, A. Kejariwal, A novel technique for long-term anomaly detection in the cloud., in: HotCloud.

[31] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016).

[32] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, arXiv preprint arXiv:1607.01759 (2016).

[33] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651 (2016).

[34] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS'13, Curran Associates Inc., USA, 2013, pp. 3111–3119.

[35] H. R. Glahn, D. A. Lowry, The use of model output statistics (MOS) in objective weather forecasting, Journal of Applied Meteorology 11 (1972) 1203–1211.

[36] S. Streaming, Spark Streaming Overview, `https://spark.apache.org/streaming/`, 2017. [Online; accessed 20-August-2017].

[37] PostgreSQL, PostgreSQL, `https://www.postgresql.org/`, 2017. [Online; accessed 20-August-2017].

[38] P. Nurmi, A. Perrels, V. Nurmi, Expected impacts and value of improvements in weather forecasting on the road transport sector, Meteorological Applications (Special Issue) 20 (2013) 217–223.

[39] A. Elisseeff, M. Pontil, et al., Leave-one-out error and stability of learning algorithms with applications, NATO science series sub series iii computer and systems sciences (2013) 111–130.

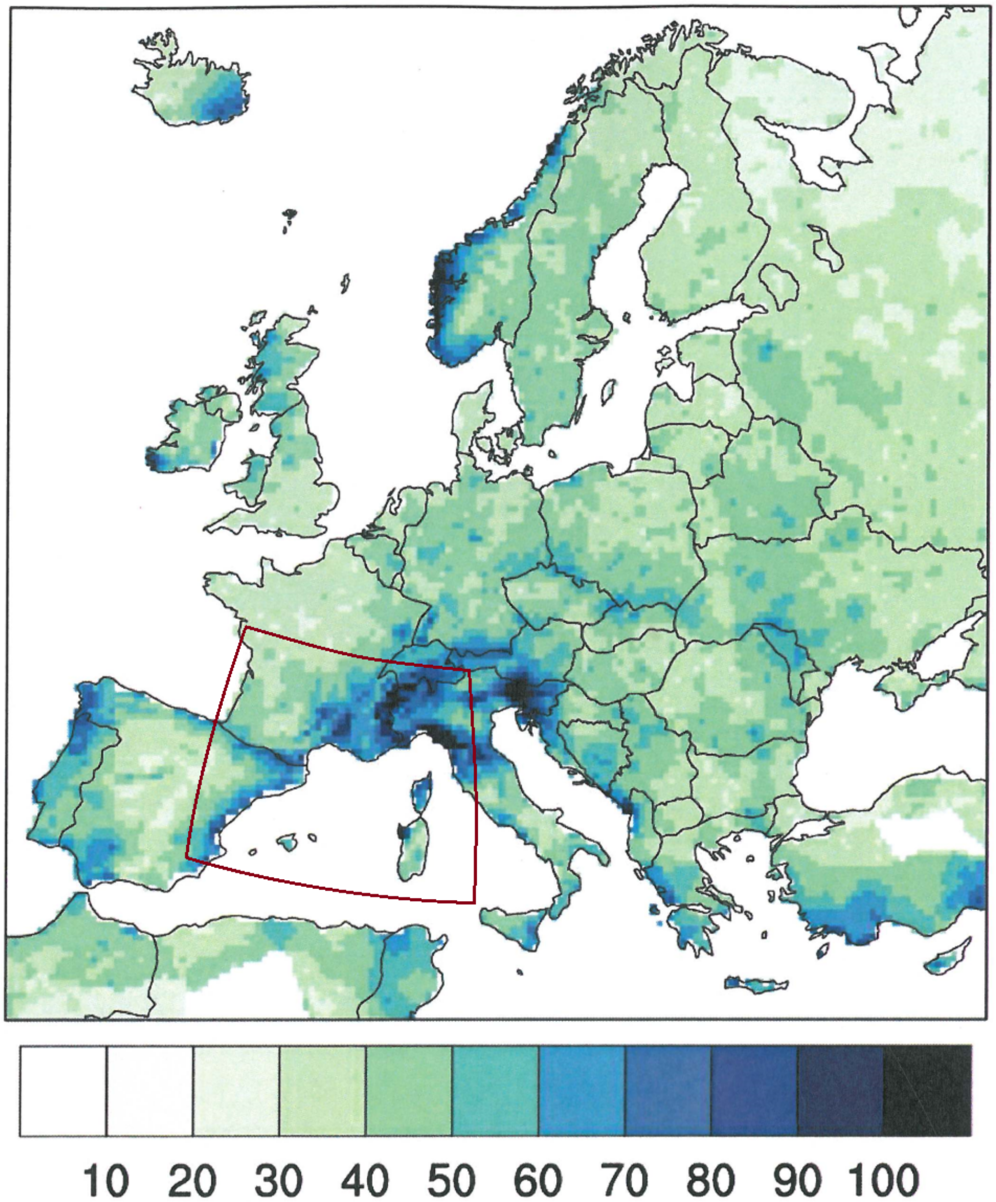

Figure 1: 10-year return level of daily precipitation (mm) according to E-OBS data set for the period 1981-2010 (Groenemeijer et al, 2016 [8]). The red box indicates the region investigated in Figure 3.

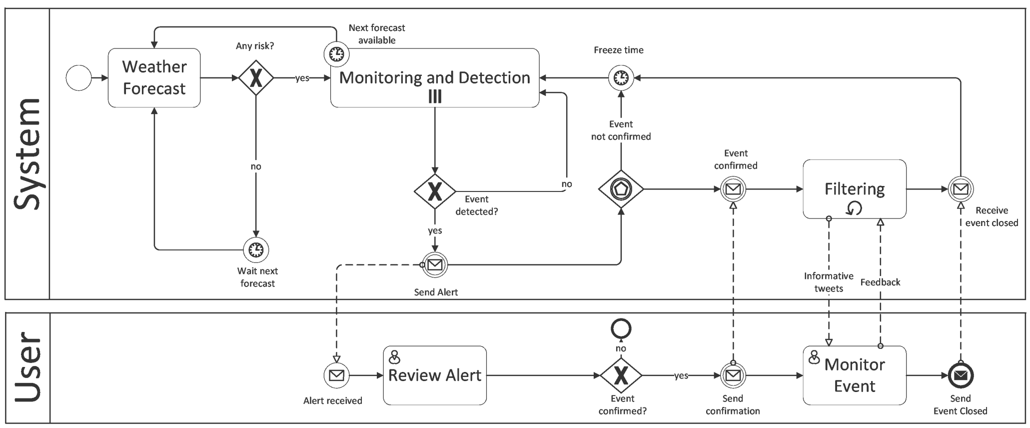

Figure 2: Flow of the proposed set of services. The diagram is realized according to the Business Process Modeling Notation (BPMN).

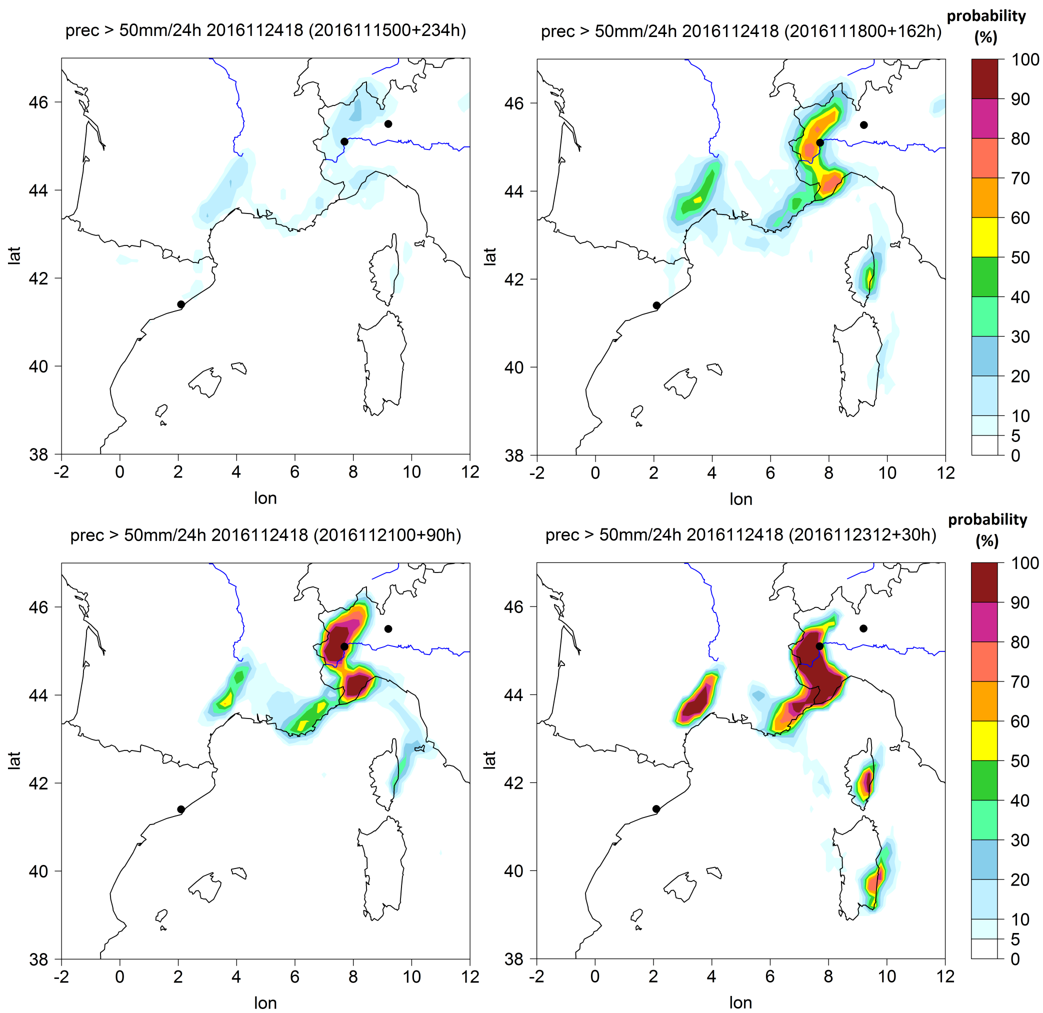

Figure 3: Probabilistic forecasts for accumulated precipitation to exceed 50 mm/24h. Every forecast is valid on 24$^{\text{th}}$ of November 2016 but they have different lead times: 234, 162, 90, and 30 hours. Analysis time and lead time is shown in parentheses. Black points mark Barcelona (leftmost one), Turin and Milan, and blue lines mark the rivers Loire (France) and Po (Italy).

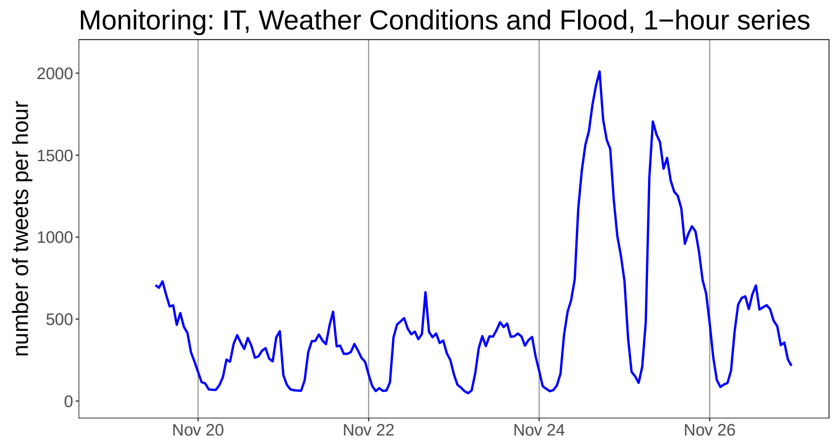

Figure 4: Tweets volume in the considered time range (from 19th to 26th of November, 2016).

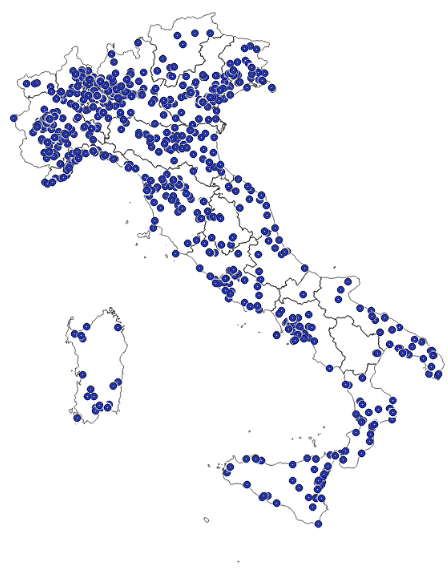

Figure 5: Tweet localization in the considered time range (from 19th to 26th of November, 2016).

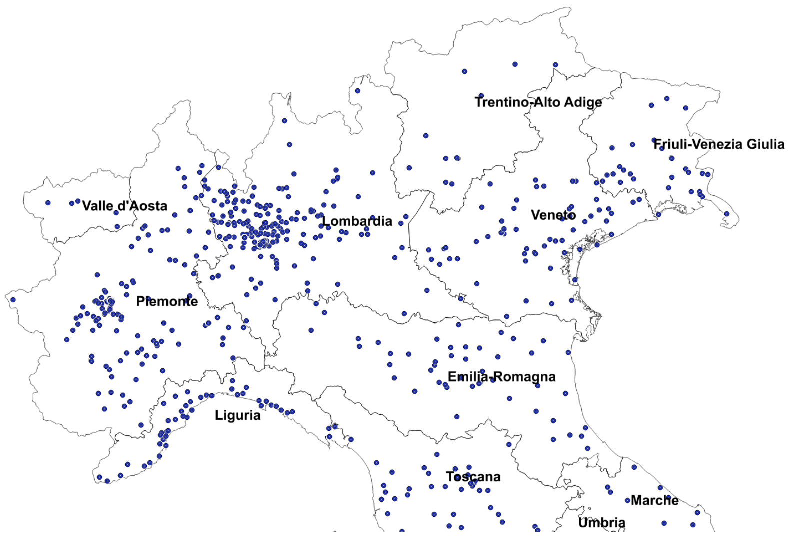

Figure 6: Tweet localization in the considered time range (from 19th to 26th of November, 2016), focusing on the affected areas.

Figure 7: Key Phrases (23 November)

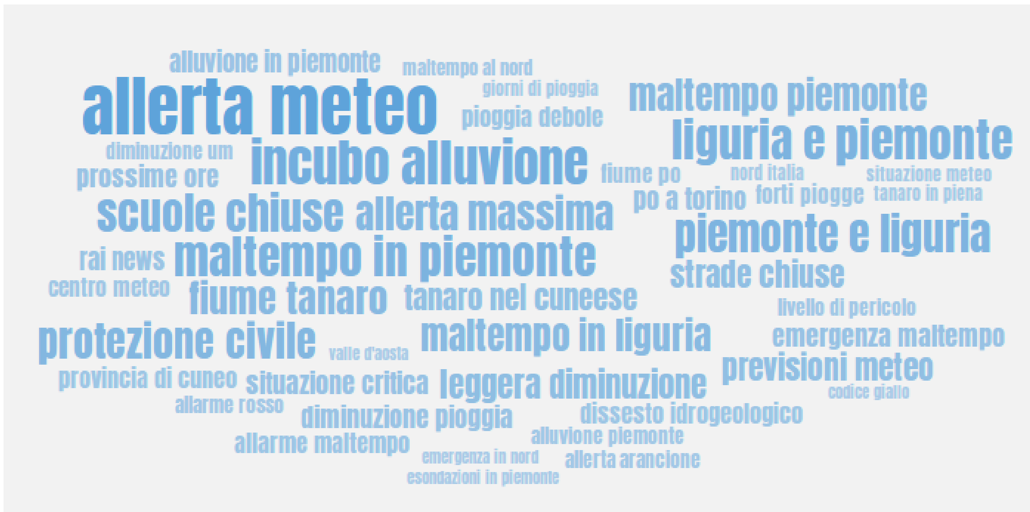

Figure 8: Key Phrases (24 November)

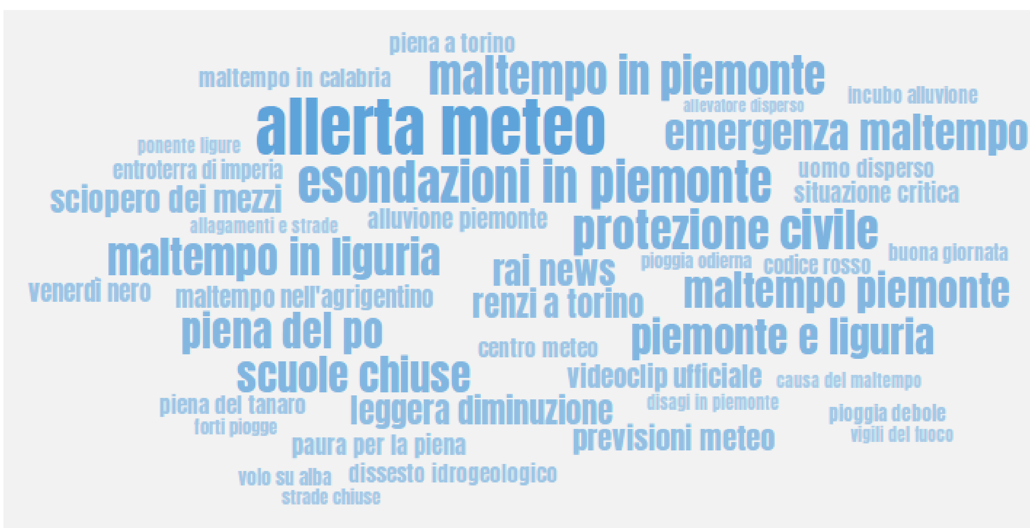

Figure 9: Key Phrases (25 November)

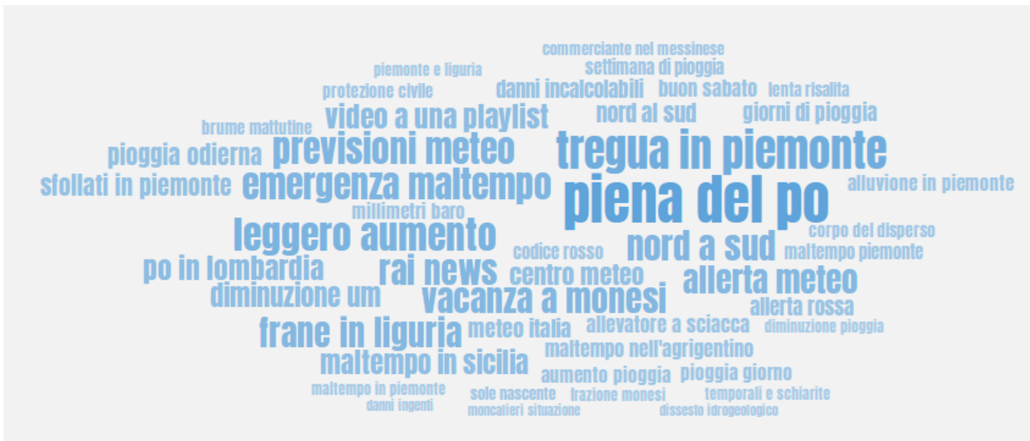

Figure 10: Key Phrases (26 November)

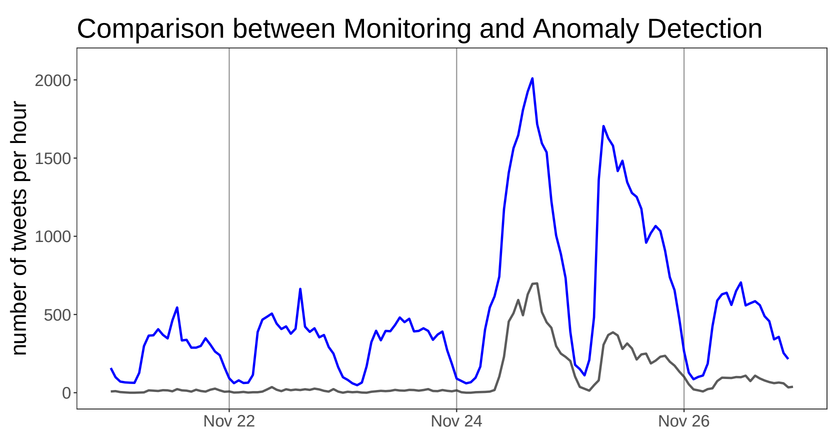

Figure 11: Comparison between Monitoring and Anomaly Detection streams for Italian, between November 21st and 27th 2016. The Monitoring stream (blue) contains Tweets related to weather conditions and Floods. The Anomaly Detection stream (gray) contains only Tweets related to Floods.

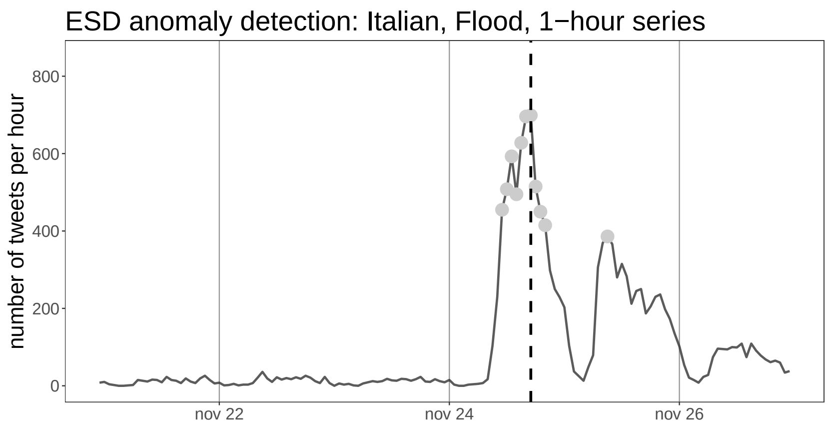

Figure 12: Results of anomaly detection using ESD test on the 1-hour series of Italian Tweets concerning Floods, between November 21st and 27th 2016. Anomalies appear as large dots. First anomaly is detected on November 24th at 11:00 local time. Dotted line marks the Event time according to Copernicus: November 24, 18:00 local time.

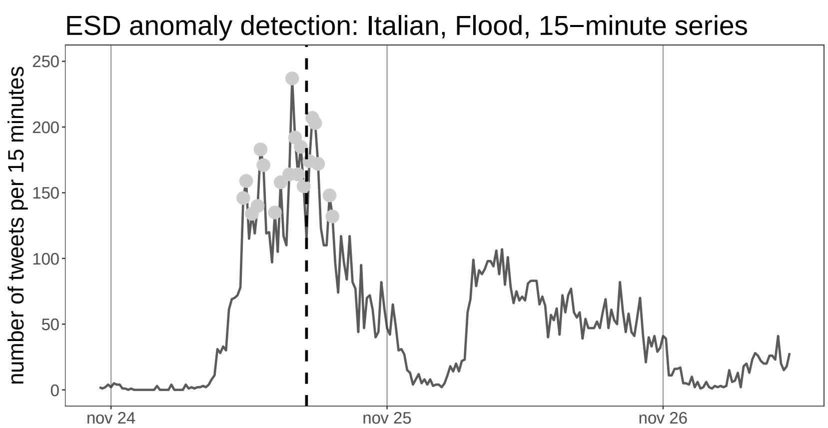

Figure 13: Results of anomaly detection using ESD test on the 15-minute series of Italian tweets concerning Floods, between November 24th and 26th 2016. Anomalies appear as large dots. First confirmation of 1-hour anomaly is obtained on November 24th at 11:30 local time. Dotted line marks the Event time according to Copernicus: November 24, 18:00 local time.

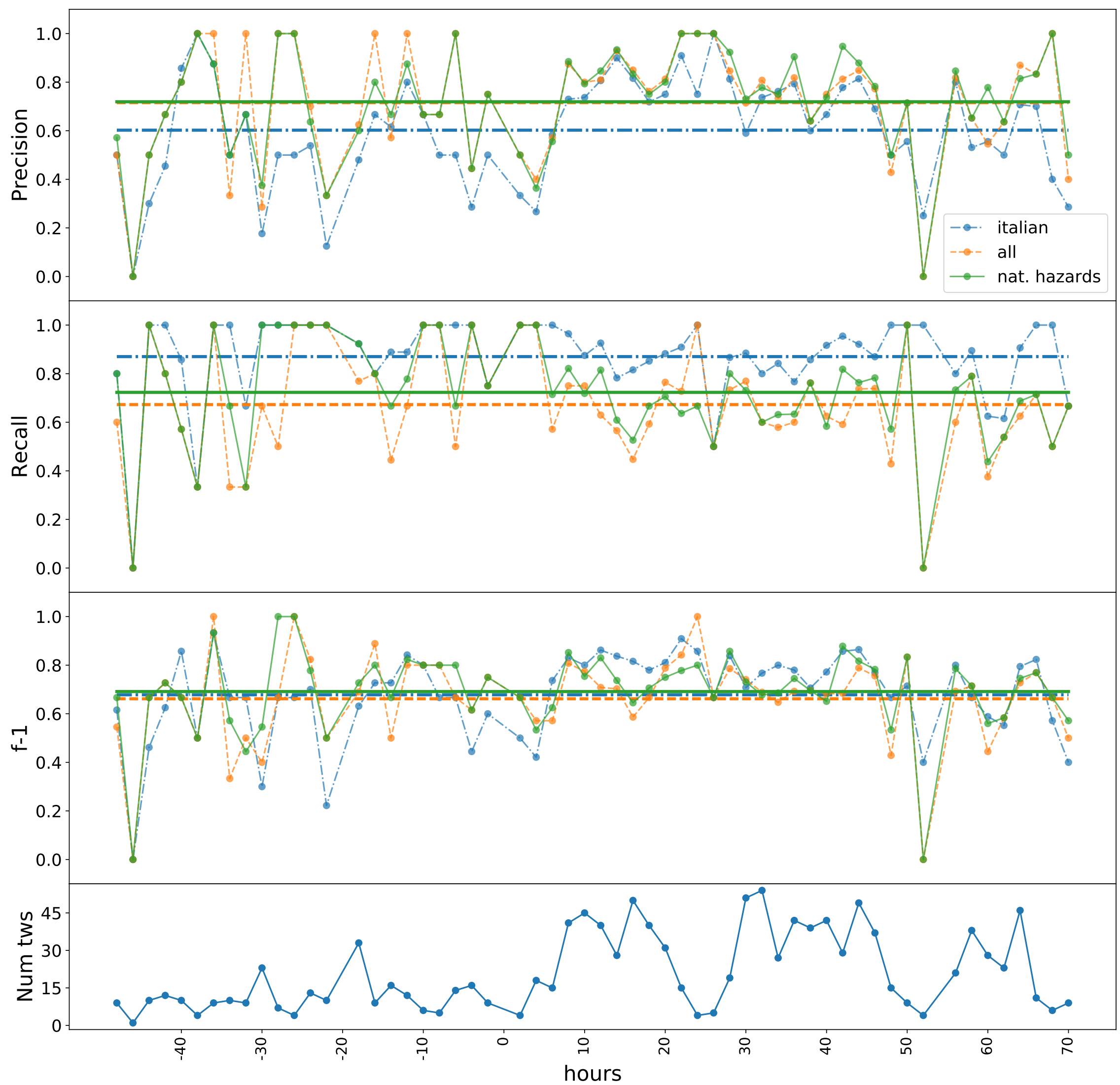

Figure 14: Performance obtained for a selection of annotated Tweets collected before and during the Piedmont flood in November 2016. The different lines refer to different datasets used for training the `fasttext` algorithm. "Italian": Tweets only in Italian are considered, "all": Tweets from all the 26 CrisisLexT26 events, "nat. hazards" Tweets related only to natural hazards events. The dip in performances at -46 and 52 hours is connected with the very low number of tweets connected with those intervals (1 and 4 tweets).