UNIVERSITÀ
DEGLI STUDI
FIRENZE

PhD in
Earth Sciences
CICLO XXXV

COORDINATOR Prof. Sandro Moretti

Analysis of community resilience during natural disasters using data mining on massive social networks exchanges

Doctoral Candidate

Dr. *Rachele Franceschini*

Supervisor

Prof. Filippo Catani

Co-Supervisor

Dr. Ascanio Rosi

Coordinator

Prof. Sandro Moretti

Years 2019/2022

Abstract

Mass media are a new and important source of information for any natural disaster, mass emergency, pandemic, economic or political event, or extreme weather event affecting one or more communities in a country. Mass media is usually the first and primary source of information about hazards for the public, providing a relatively high temporal and spatial resolution. Various studies have shown that mass media have a quick degree of observation and publication of the event in a relatively short period. The use of data mining techniques is advancing in different ways. The news publication about a natural disaster on newspaper or crowdsourcing platforms allows a faster observation, survey, and classification of these phenomena. This source of information allows continuous feedback from the real world, and news concerning natural events can be rapidly collected.

The major goal of this research is to show how useful and capable social media is for detecting events in places without actual sensors that could immediately identify a natural hazard. For the entire Italian area, many analytical techniques have been used to determine the spatial and chronological distribution of newspaper articles about floods and in particular on landslide events. This analysis made it possible to identify the areas with greater user interaction concerning a natural event. User interaction can identify active social behaviour, inclined to information (transmitting and/or receiving it) and consequently resilient to the event. The procedures that can be taken to manage the data and show what can be derived are discussed below in great detail. In the first step, news of landslides and floods was analysed using as source the Multi-risk Information Gateway or MIG platform, which collected articles about natural events (landslides and floods) at the national scale from Google News. The newspaper articles about landslides and floods in Italy are automatically collected by an existing data mining algorithm based on a semantic engine and archived within multi-risk information gateway platform. In total, 32.525 landslide news items and 34.560 flood news items were collected. The datasets are classified into classes based on the thematic, temporal and spatial relevance of the news. This classification makes it possible to outline the temporal and spatial distribution of the events, their media impact and also to outline a hazard map on a regional and provincial scale. There are three classes in total, class 1 with high temporal and spatial reliability, class 2 has a medium temporal and spatial resolution, while class 3 identifies incorrect news. These dataset feature inventories of events from 2010 to 2019. The integration of natural hazard information and social media data could improve warning systems to enhance the awareness of disaster managers and citizens about emergency events. Currently, few studies have been produced on the combination of social media data and traditional sensors. This gap indicates that it is unclear how their integration can effectively provide emergency managers with appropriate knowledge. A landslide and flood inventory derived from social media was

used as a base proxy to correlate rainfall data and impacts of events in terms of victims (POLARIS) and earmarked funds (ReNDIS). However, more emphasis was placed on the problems of landslide events by considering maps of landslide hazard percentages, population at risk and buildings at risk (from ISPRA). This can attempt to show how social media, combined with other sources, can assist government authorities with a better knowledge of the hazard of a territory. These data have been used to identify the areas and the periods most affected by natural hazards. In another hand, it was possible to outline the resilience of communities/regions considering the number of published news with respect to natural hazards. Landslide news, with its associated media impact and number of victims, is mainly concentrated in the northern regions and Campania and Sicilia. A similar trend was found in the distribution of rain events. Conversely, the distribution of funds is more concentrated in the South rather than in Northern Italy. This trend was justified by the large percentages of landslide hazard areas and buildings at risk that characterise them. Whereas, the regions most affected by floods were mainly central and central-southern Italy. This trend agrees with the distribution of rainfall, the number of victims and the funds allocated for soil protection. Considering the temporal distribution, in general, both landslide and flood events have increased since 2015, in contrast to the rainfall data and allocated funds, while casualties remain stable.

According to the results, data mining is helpful for creating databases where the day and the approximate location (municipality) of the possible events are known. This database can be used for proper land use or risk mitigation planning since the most event-prone municipalities can be defined.

In the second step of this work, a new data mining technique in Twitter was applied using appropriate keywords extracted from newspaper headlines. Several techniques have been developed for data mining in social media for many natural events, but they have rarely been applied to the automatic extraction of landslide events. Currently, several systems to set up landslide inventories exist, although they rarely rely on automated or real-time updates. For these reasons, this work focused on landslide events. This makes it possible to fill the gap in the literature with respect to landslide events. One script was set to obtain the database from Twitter. Tweets were collected through the Twitter academic API2 with an academic licence. The Twitter dataset comprises various slots with different temporal distributions. The main purpose is not to recreate the same inventory of landslide phenomena as with newspaper reports but rather to apply classification techniques. Hence, the dataset from Twitter is considered neither complete nor exhaustive for the 2011-2019 period. The dataset features by 13.349 data, it was classified manually, providing a solid base for applying deep learning. Exploring the dataset, some case studies were analysed. An interesting case involved the landslide event of 24/11/19 in Liguria. For this event, possible alert maps were created at municipal scale and by alert zone considering the count of tweets. The alert maps demonstrated how the data, from social media, at the municipal scale is still comprehensive in the civil protection phases. In addition, the timing of user

interaction with the event was demonstrated compared to the slower publication of newspaper articles. As a result of these analyses, a possible contribution to the implementation of specific guidelines for communication and alerting about natural events such as landslides was proposed. Creating a simple, uniform language can facilitate communication between decision-makers and citizens, and also between decision-makers and data analysts.

In the third step of this project, it was applied deep learning model for classifying the dataset on the basis of landslide information. A script with transformer architecture and the BERT method was created to classify the data. "**Bert For Information on Landslide Events**", or BEFILE, allows the classification of text into two classes (0 and 1) based on landslide information in the Italian language. The Italian-language classified dataset for landslide events fills the gap in analysing natural events using Twitter, which has not yet been exploited to a great extent for landslide events. BEFILE makes possible the detection of landslide events within tweets and brings state-of-the-art integration in NLP technology of text classification. At the same time, several problems may arise due to the nature of big social media data analysis and some limitations of this research. These problems should not be ignored when translating the research results into practice. However, it was demonstrated that Twitter can be utilized as a source of rapid information and detection for landslide events. A possible contribution about implementation of specific communication and warning guidelines with respect to natural events such as landslides has been proposed. Creating a simple homogeneous language can available the communication between decision-makers and citizens, but also decision-makers and data analysis-makers. Moreover, from a practical perspective, this study provides useful perspectives for decision-makers to consider when using social media as an additional information resource for rapid damage assessment.

Riassunto

I mass media sono una nuova e importante sorgente di informazione per i disastri naturali, emergenze di massa, eventi politici o eventi meteo estremi che coinvolgono ona o più comunità di un paese, fornendo un'alta risoluzione temporale e spaziale. L'uso di tecniche di data mining, applicato a diverse sorgenti (Google News o piattaforme social) si sta diffondendo in diversi campi. La pubblicazione di una notizia o di un post da un utente all'interno di una piattaforma crowdsourcing permette una veloce osservazione e classificazione del fenomeno. L'accesso a queste informazioni permette di ottenere un feedback continuo, dal mondo reale. Tale accesso aumenta notevolmente la risposta degli organi competenti di Protezione Civile ad assistere all'emergenza.

In questo lavoro diverse fasi e analisi sono susseguite per dimostrare l'utilità e l'efficacia di utilizzare anche i dati dai social media per rilevare eventi naturali e determinare la resilienza di un'area.

Nella prima fase sono state analizzate le notizie di frane e alluvioni utilizzando la piattaforma Multi risk Information Gateway o MIG. Questa piattaforma è stata sviluppata all'interno del Dipartimento di Scienze della Terra dell'Università degli Studi di Firenze. Un algoritmo semantico filtra, raccoglie e cataloga le notizie provenienti da Google News. In totale sono stati collezionati 32.525 notizie di frana e 34.560 notizie di alluvione. I datasets sono classificati in classi sulla base della rilevanza tematica, temporale e spaziale della notizia. Le classi in totale sono 3, la classe 1 delinea un'alta affidabilità temporale e spaziale, la classe 2 presenta una media risoluzione, mentre la classe 3 identifica le notizie errate. Per distinguere i vari prodotti generati dalla classificazione, una nuova nomenclatura è stata creata delineando gli eventi, la pericolosità da frana in un'area e l'impatto mediatico o l'impatto dell'evento. Tale classificazione, quindi, permette di determinare i relativi prodotti sotto forma di distribuzione temporale e spaziale. Nella distribuzione temporale sono stati considerati la distribuzione annuale, mensile e giornaliera, mentre sono state studiate tre risoluzioni spaziali: regionale, provinciale e di zona di allerta. In conclusione, datasets così classificato vanno a formare degli inventari rispettivamente di frane e alluvioni per tutto il territorio italiano dal 2010 al 2019. In special modo, le analisi di distribuzione spaziale del dato ha permesso di stimare la resilienza in funzione degli articoli pubblicati per ogni regione. Per convalidare tale analisi, l'attenzione successiva si è rivolta ad integrare i dati dai social media con ulteriori informazioni sui rischi naturali provenienti dai sensori tradizionali o da altri fonti di dati disponibili. I precedenti inventari sono stati utilizzati come proxy di base per correlare differenti dati: pluviometrici ed effetti dell'evento in termini di perdite di vite umane (POLARIS) ed economiche (ReNDIS). Una maggior analisi è stata posta sulle problematiche degli eventi di frana considerando anche le mappe delle percentuali di pericolosità da frana, popolazione a rischio e edifici a rischio (da ISPRA).

Le notizie di frana, con il relativo impatto mediatico e numero di vittime si concentrano principalmente nelle regioni settentrionali e in Campania e in Sicilia. Tendenza simile è stata riscontrata nella distribuzione degli eventi di pioggia. Al contrario, la distribuzione dei fondi si concentra maggiormente al Sud rispetto al Nord Italia. Tale trend è stato giustificato dalle notevoli percentuali di aree a pericolosità da frana ed edifici a rischio che li caratterizzano. Mentre, le regioni più colpite dalle alluvioni sono state soprattutto l'Italia centrale e centro-meridionale. Questa tendenza concorda con la distribuzione delle precipitazioni, il numero di vittime e dei fondi stanziati per la protezione del suolo. Considerando la distribuzione temporale, in generale sia gli eventi franosi che alluvionali aumentano dal 2015, in contrasto con i dati di pioggia e i fondi stanziati, mentre le vittime rimangono costanti.

Data l'attuale letteratura sul data mining per gli eventi alluvionali e l'assenza di studi sugli eventi franosi, le analisi successive si sono concentrate su quest' ultimi. Questo ha permesso di approfondire e analizzare un tema non realmente affrontato nelle analisi dei dati dalle piattaforme di crowdsourcing. Nella seconda fase, di questo progetto, è stato utilizzato Twitter come fonte di dati. La tecnica di data mining, finora applicata alle notizie dei giornali, è stata applicata alla piattaforma crowdsourcing usando un script e parole chiavi appropriate. All'interno della pagina per sviluppatori di Twitter è stato creato un progetto con accesso accademico. Il progetto è supportato da una applicazione appositamente creata per ottenere le credenziali, le quali consentono di estrarre il dato dalla piattaforma utilizzando Python. Lo script di estrazione preesistente è stato settato considerando 5 mirate parole chiavi afferenti all'evento frana. Le parole chiavi sono state scelte sulla base di un'analisi semantica all'interno dei titoli di giornali afferenti agli eventi di frana. Mentre i periodi temporali di estrazione sono stati basati sulla distribuzione temporale delle notizie. Nove slots sono stati estratti con differente risoluzione temporale, per un totale di 13.350 tweets. Il dataset è stato classificato manualmente considerando la rilevanza e la presenza di coordinate specifiche sul testo. Esplorando il set di dati, sono stati analizzati alcuni casi di studio. Caso interessante è riguardato l'evento di frana del 24/11/19 in Liguria. Per questo evento sono state create delle possibili mappe di allerta a scala comunale e per zona di allerta considerando il conteggio dei tweets. Le mappe di allerta hanno dimostrato come i dati, dai social media, a scala comunale siano comunque esaustivi nelle fasi di protezione civile. Inoltre, è stato dimostrato la tempistica di interazione degli utenti all'evento rispetto alla più lenta pubblicazione di articoli di giornale. A seguito di queste analisi è stato proposto un possibile contributo per l'implementazione di linee guida specifiche per la comunicazione e l'allerta in relazione a eventi naturali come le frane. La creazione di un linguaggio semplice e omogeneo può favorire la comunicazione tra decisori e cittadini, ma anche tra decisori e addetti all'analisi dei dati.

Tuttavia, occorre tenere conto di alcune limitazioni: le parole chiavi utilizzate possono essere non esaustive e di conseguenza il dataset può essere non completo, inoltre, Twitter limita il numero di estrazioni per unità temporale; infine, la mancanza spesso di geolocalizzazione del dato.

Il terzo step di questo lavoro è stato caratterizzato dalla applicazione di tecniche di deep learning. Il dataset classificato ha fornito una solida base per l'applicazione di deep learning supervisionato. Inoltre, il dataset classificato in lingua italiana per gli eventi franosi colma l'attuale lacuna nell'analisi degli eventi naturali. Infatti, Twitter allo stato attuale non è ancora stato sfruttato per questa tipologia di evento. Uno script è stato creato per la classificazione del testo, utilizzando un'architettura a trasformatori con il metodo BERT. "**Bert For Information on Landslide Events**" o BEFILE permette di classificare il testo in due classi (0 e 1) in base alle informazioni sulle frane in lingua italiana. Questa analisi porta a un notevole avanzamento del classificatore BERT, che finora era stato utilizzato spesso per analizzare dati in lingua inglese in diversi settori. BEFILE senza preelaborazione ha mostrato risultati significativi di accuratezza, pari al 96% e un'AUC di 0,95; posizionandosi tra l'implementazione di modelli con CNN. BEFILE ha mostrato risultati promettenti nella classificazione e quindi nell'individuazione di informazioni su eventi franosi.

Nonostante i limiti noti e dimostrati dei dati dei social media, questo studio conferma che informazioni rilevanti e statisticamente significative sulla pericolosità delle frane e delle alluvioni possono essere ottenute attraverso il data-mining dei social network durante le emergenze. Tali dati, opportunamente filtrati e classificati, possono essere di notevole aiuto per aumentare la nostra attuale capacità di calibrare e validare i modelli di allerta precoce, con particolare riferimento alle aree con scarsità di dati. Inoltre, alcune valutazioni possono rappresentare uno strumento utile per comprendere e valutare l'impatto dei disastri naturali, nonché per pianificare le migliori strategie di riduzione del rischio su scala regionale o nazionale.

Index

1 INTRODUCTION

The vulnerability of the population and the effects of natural hazards depend on the geological and geomorphological settings of an area even on the level of socioeconomic development. A recent report from US National Academy of Science recognizes a growing threat in relatively wealthy municipalities in the United States, with higher risk to populations with less protection from insurance or social safety net (National Academies of Sciences & Medicine, 2019). Production growth and accelerated urbanization, as well as the concentration of the population and enterprises in hazardous zones, are the main factors of the increase in risk (Crozier, 2010; World Conference on Disaster Risk Reduction, 2015; Porfiriev, 2016). Climate change is expected to make these conditions even worse (Kundzewicz et al., 2014; Hirabayashi et al., 2013). As result of global climate changes, extreme weather events are predicted to increase in frequency and severity at regional and local scales (IPCC, 2012). Kiely (1999) affirmed that the rainfalls with heavy intensity rate are increasing and the probability of the return period of relevant events are shortening from 30 to 10 years. Consequently, the frequency and severity of natural disasters on the economy are growing. Estimating the economic impact and the human losses outline as consequences of natural disasters is a necessary tool to evaluate and develop measures for risk reduction. Several research projects have examined the economic impact of specific disaster events (Cavallo and Noy, 2009). The wealthier the country affected, the greater the economic losses from natural disasters (Neumayer et al., 2014). The natural hazard propensity and the possible involvement of urbanized areas are issues faced by governments and private actors in undertaking measures that will prevent, or at least mitigate economic losses. Countries of larger economic size will have more wealth potentially destroyable and are therefore expected to experience larger losses. Also, disaster prevention and damage mitigation measures are costly and both private actors and governments can more easily finance the in richer countries rather than in low social capital countries (Neumayer et al., 2014). Hurricane Katrina was the costliest natural disaster ever with an estimate economic loss between 82 billion (Knabb et al., 2005) and 150 billion US$ (Burton and Hicks, 2005), and 986 victims (Brunkard et al., 2008). In Italy, in October 2014, one flash flood caused several landslides and mud flows in Genova, causing 1 dead, 300 million euros of damage and 250 people homeless (Faccini et al., 2015; Paliaga et al., 2020).

Landslides in Italy are the most frequent and diffuse natural hazards causing the greatest number of fatalities and damage to urban areas and infrastructure (Forli and Guida, 2009; Campobasso et al., 2013). Landslides are an illustrative example of multi-hazard, which can be caused by earthquakes, rainfalls and human activity among other reasons. Detection of landslides presents a significant challenge since there are no physical sensors that would detect landslide directly (Musaev et al., 2015). The retrieval of data, using specific data mining algorithms, from technical reports and/or newspapers can further extend the exploitable data. Mass media are a new and important source of information for any natural disaster, mass emergency, pandemic, economic or political event, extreme weather event affecting one or more communities in a country. Various studies have shown that mass media have a quick degree of observation and publication of the event in a relatively short time span. The use of social media in detecting natural hazard has shown promising results (Holderness et al. 2015; Wang et al., 2018). The joint analysis of data from different social media can help to capture disaster situations with a relatively high temporal and spatial resolution to map different events, such as landslides, across various locations (Fan et al., 2018; Rachunok et al., 2019; Saltelli et al., 2020). Currently, systems using automated or real-time updates are still uncommon and only used for some types of natural hazards (Battistini et al. 2013, Battistini et al. 2017; Calvello and Pecoraro 2018), mainly earthquakes, floods and wildfire, while creating a complete and updated database is more difficult for landslide (Galli et al., 2008; Santangelo et al., 2010). Only a small part of the data is used for database creation. Landslide events are completely missing from this list. Battistini et al., (2013, 2017) and Kreuzer et al., (2020) created systems for automatic real-time updating of landslide inventories using data mining techniques within newspaper articles.

Social media data have often served as proxies for a variable of interest and correlated with conventional data sources, such as physical sensors and survey data (De Andrade et al., 2021). Although, currently, few studies on the combination of social media and other data sources have been produced. It thus remains unclear how social media data can be effectively integrated with hazard-monitoring data to provide emergency managers with appropriate info for better land use planning and early warning support (Shoyama et al., 2021). Filling the gap would allow us to also outline the resilience of the population using social media as source and validate it to other available datasets that describe the territory (hazard maps, rainfall distribution and earmarked funds). Some scholars have demonstrated as social media can be an assessment component of resilience developed by individuals (Zhang & Shay, 2019). Current literature presents hundreds of definitions of the term resilience (Aburn et al., 2016; Reich et al., 2010). It is usually defined as the capacities or processes that assist the entity to prepared for upcoming disasters or cope with efficiently (by responding and withstanding) to emergency and to recover and bounce back from disasters and change (Leykin et al., 2018). Dufty et al., (2012) and Leykin et al., (2018) propose social media as base to evaluate build community resilience

to disasters through risk reduction, emergency management, and post-hazard development. Such analysis confirmed that communities with higher resilience capacity, which are characterized by better social–environmental conditions, tend to have higher social media or crowdsourcing platforms use (Wang et al., 2021). Reports about stressful events in social media news sites trigger various appraisals among social media users. Some of these appraisals are translated into textual expressions and offer a unique opportunity to observe how the public "digests" and copes with the changing reality (Leykin et al., 2018). These results imply that social media, such as Twitter, use during disasters could be improved to increase the resilience of affected communities (Wang et al., 2021).

The main aim of this work is to demonstrate the utility and capability of social media to detect events in areas without physical sensors and in consequence based on publications number defining the region most resilient. Different steps of analyses have been applied to define the spatial and temporal distribution of events considering newspaper articles for whole Italian territory. Below many steps are described for demonstrating how it is possible to manage the data and what it is possible to derive.

In the first section of this work was to fill the gap between social media and traditional sensors. Newspapers from Google News were collected by Semantic Engine to Classify and Geotagging News (SECAGN) system developed by Battistini et al., 2013,2017 were exploited and studied. In particular, landslide and flood events from 2010 to 2019 were considered for the entire Italian territory. The disadvantage created by this type of data is heterogeneity. Many textual data are congruent with the event, but some texts with incorrect word associations manage to evade the filtering system. In this case, a manual classification was carried out to further facilitate the analysis. The dataset will later form the basis for multiple statistical. A landslide and flood inventory derived from social media was used as a base proxy to correlate rainfall data and impacts of landslides in an attempt to show how social media in combination with other sources can be utilized to assist government authorities with a better knowledge of the landslide hazard of a territory. Such analysis, further, allowed to outline the resilience at regional scale, considering the number of articles published respect to natural event. Methods and results of this part have been illustrated respectively in Chapter 3.1, 3.2 and Chapter 4.1.

In the second section of this work, a new data mining technique within Twitter has been applied. Crowdsourcing platforms such as Twitter, Instagram, Facebook or YouTube are widely used to detect different types of events. In America, as in Asia automatic data mining systems are applied for earthquake, flood, hurricane and fire events. Specifically, Twitter is an excellent resource for event detection. People share opinions and information about the situation. Especially with Twitter being used as a medium of communication every day, the amount of various information about different events that can be found is overwhelming (Madichetty and Sridevi, 2020). Having easy access to tweets

coming from people would afford new possibilities for emergency response, such as a contribution to the real-time assessment of impacts, criticalities and needs. In this project, appropriate keywords have been extracted from newspaper headlines. Given the present literature on data mining for flood events and the absence of studies on landslide events, the analysis focused on the latter events to deepen and analyse a topic not truly addressed in social media analyses and crowdsourcing platforms. This makes it possible to fill the gap in the literature with respect to this hazard event. Over 13.000 data were extracted within Twitter using keywords about landslides. The dataset was classified manually, using two classes (0 and 1), based on subject relevance. Exploring the dataset, some case studies have been analysed. Based on tweet count possible Alert system maps were created. These demonstrated how the data on a municipal scale is in any case exhaustive in the civil protection phases. However, the dataset can be considered not exhaustive mainly for two reasons and in different fields: i) the limitation of Twitter during extractions and ii) the lack of geo-localization of data sometimes not provided. However, it was demonstrated as also Twitter can be utilized as a source of rapid information and detection for landslide events. From a practical perspective, this study can provide useful perspectives for decision-makers to consider when using social media as an additional information resource for rapid damage assessment. At the end, it was proposed a possible contribution about to the implementation of specific communication and warning guidelines with respect to natural events such as landslides. Creating a simple homogeneous language can available communication between decision-makers and citizens, but also decision-makers and data analysis-makers. Methods and results of this section have been illustrated respectively in Chapter 3.3 and Chapter 4.2.

Using tweets has become one of the most important tools for natural language processing (NLP) tasks. In the third section of this project, the main aim is to obtain an automatic classification based on information about landslide events. Disaster tweet classification study can be considered a natural language processing task. The dataset from Twitter, classified manually, provided a solid base for applying deep learning. The use of a deep learning model has started to become more common for natural language processing tasks. Moreover, the Italian-language classified dataset for landslide events fills the present gap in analysing natural events using Twitter, not yet exploited to a great extent for landslide events. The transformer architecture has been chosen for text classification within deep learning with the method BERT. “Bert For Information on Landslide Events” or BEFILE is the script created to classify text into two classes (0 and 1) based on landslide information in Italian language. This analysis leads to a considerable advancement of the BERT classifier, which until now was very often only used for a variety of analyses in English. BEFILE without preprocessing showed important values of accuracy, equal to 96% and AUC of 0,95; located between implementing models with CNN.

Some validations were applied considering the distribution of news and tweets during the events. Methods and results of this section have been illustrated respectively in Chapter 3.4 and Chapter 4.3.

1.1 Mass media for natural disasters

Mass media is generally the first and primary source of information about hazards for the public (Fischer, 1994). The use of social media in detecting natural hazards has shown promising results (Holderness et al. 2015; Wang et al. 2018). Studies indicate that social sensors in terms of social media report a natural disaster much faster than do observatories (Goswami et al., 2018). Such characteristics provide a unique opportunity to capture disaster situations with a relatively high temporal and spatial resolution. Furthermore, different events can be mapped across various locations (Fan et al., 2018; Rachunok et al., 2019). Systems using automated or real-time updates are still uncommon and only utilized for some types of natural hazards (Battistini et al., 2013, Battistini et al., 2017), mainly earthquakes, floods, hurricanes and wildfires. Social networking sites have multiple roles. The creation of databases about natural disasters is an application undertaken for hurricanes (Miles et al., 2007) or flood (Du et al., 2015) events, forecasting disasters (Huang et al.,2010), and focusing on warnings (Acar & Muraki, 2011) and postcrisis activities (Olteanu et al.,2015). Creating a complete and updated database is more difficult for landslides (Galli et al., 2008, Santangelo et al., 2010). Landslide research chiefly relies on landslide inventories for a multitude of spatial, temporal or process analyses (Van Den Eeckhaut and Hervás, 2011; Kirschbaum et al., 2015; Klose et al., 2015). The forecasting and displacement monitoring of landslides are being increasingly characterized as a problem of “big data”. Different data sources can be employed to support decision-making: satellites (Soeters and Van Westen, 1996; McKean and Roering, 2003; Lu et al., 2012; Bianchini et al., 2018; Montalti et al., 2019; Solari et al., 2020; Confuorto et al., 2021; Nava et al., 2022), rainfall gauges (Lagomarsino et al.,2013; Segoni et al., 2018, Rosi et al., 2021) and hydrological networks (Horita et al., 2017). The retrieval of data from technical reports and/or newspapers, using specific data mining algorithms, further extends exploitable data. The methodology of Battistini et al., (2013, 2017) and Kreuzer et al., (2020) allows us to update the landslide inventory in near real-time using the data mining technique of online newspaper articles.

Most of the empirical literature that employs social media data has focused on investigating the relationships between social media data and real-world phenomena. This approach involves extracting aggregated, thematic, spatiotemporal patterns from social media activity. Social media data have often served as proxies for a variable of interest and correlated with conventional data sources, such as physical sensors and survey data (De Andrade et al., 2021). Baranowski et al., 2020 and Fitriany et al.,

2021 combined social media information about floods and forest fires with rainfall data and satellite and wind velocity data. The authors have demonstrated how an additional source of information can be utilized for near-real-time forecasting. However, currently, only a few studies on the combination of social media and other data sources have been produced. It thus remains unclear how social media data can be effectively integrated with hazard-monitoring data to provide emergency managers with appropriate early warnings (Shoyama et al., 2021).

Crowdsourcing platform - Twitter

Social media with crowdsourcing functions, such as Twitter, Facebook or Instagram, are increasingly being utilized as sources in mainstream news coverage. Crowdsourcing platforms have become an indispensable part of people's everyday lives and a powerful tool of communication during emergency situations, such as during natural disasters. Many research papers about the use of social platforms in difficult circumstances have been published (Dragović et al., 2019). Among the different crowdsourcing platforms, Twitter has been extensively used for detecting natural disasters. The reason for this use of Twitter is that information appears promptly and can be effectively accessed and processed (Alam et al., 2019). Twitter had more than 321 million active users in 2020 (TIZ, 2020). Tweet features such as short messages (maximum of 280 characters) published in real-time, the ability to attach pictures and to share GPS geolocation, and the provision of a free streaming application programming interface (API) make it possible to automate monitoring tasks (Fayjaloun et al., 2021) for different events (elections, humanitarian crises such as pandemics or wars, natural disasters, etc.). People post situation-sensitive information on social media related to what they are experiencing, witnessing, and/or hearing from other sources (Hughes & Palen, 2009). With an average rate of 0,85%-3% of tweets being geo-tagged, approximately 7.000.000 geo-tagged tweets are posted per day (Huang, Li & Shan, 2018). Researchers have observed a strong and immediate spread of tweets when a significant event happens (Comunello et al., 2016; Kryvasheyeu et al., 2016). Large crises often generate an explosion of social media activity. The earliest known cases of people using the microblogging service Twitter in an emergency occurred during severe wildfires near San Diego, California (United States) in 2007 (Imran et al., 2015). Among the largest documented peaks of tweets per minute observed during disasters are 20.000 tweets per minute during Hurricane Sandy in 2012 (United State) (Castillo, 2016), approximately 17.000 tweets per minute during the Notre Dame fire in 2019 (Kozlowski et al., 2020), almost 13.000 tweets per minute immediately after the Californian Ridgecrest earthquake of July 6, 2019 (BRGM dataset), and more than 150.000 tweets in the first 48 h after the Mw 6,2 Amatrice earthquake (Italy, August 2016) (Francalanci et al., 2017). The data flow sent during the event is not constant but experienced drastic variations.

Twitter is now considered a social sensor for natural hazards by allowing shared access to live data streams. Social networking sites have multiple roles. These roles are significant in the preparation phase for a natural disaster (Table 1), during the disaster (Table 2) and after the event (Houston J.B. et al., 2015; Kim J. et al., 2018) (Table 3). In the predisaster phase or preparation phase, users on social networking sites can be alerted by certain organizations about natural disaster probability in the endangered area.

Natural disasters	year	City/region/country	Social media	Authors
Hurricane Sandy	2012	United States	Facebook, Twitter, YouTube	Bernier et al., 2013
Floods	2010/11	Queensland (Australia)	Facebook, Twitter, YouTube	Ehnis et al., 2012; Magro, 2012
Mount Merapi Eruption; Mentawai Earthquake and Tsunami; Singkil earthquake; Simeulue earthquake	2010/11/12	Japan	Twitter	Chatfield et al., 2013; Nugroho, 2011

Table 1: Case studies from the predisaster phase (Dragović et al., 2019).

The second phase, during a natural disaster, is often the most important. At that moment, help is needed for people who are endangered, and to succeed in this phase, it is important to support the spread of information (Dragović et al., 2019). Social media provides an innovative way to observe human attitudes and responses, especially during disasters.

Natural disasters	year	City/region/country	Social media	Authors
Fire	2007	San Diego, California	Twitter	Fraustino et al., 2012; Mills et al.,2009
Blizzard	2010	Bornholm, Denmark	Facebook	Birkbak et al., 2012
Earthquake and tsunami	2011	Tohoku region (Japan)	Twitter	Acar et al., 2011; Reuter C. et al., 2018
Hurricane	2012	East Coast, United States	Twitter, Instagram, Twitter	Fraustino et al., 2012; Huffington Post, 2012; Mashable, 2012; Kryvasheyeu et al., 2016
Bushfire	2013	Tasmania	Facebook	Irons et al., 2014
Tornado	2013	Moore, Oklahoma	Twitter	Blanford et al., 2014
Typhoon	2013	Philippines	Twitter	Mav Social, 2013
Snowstorm	2015	North America	Twitter	Teodorescu H.N., 2015
Earthquakes	2016	Rieti, Italy Vrancea, Romania	Twitter	Pirnau, 2017
Flood	2016	Louisiana	Facebook, Twitter	Kim et al., 2018; CNN, 2016

Table 2: Case studies from the response phase (Dragović et al., 2019).

In the recovery phase after a natural disaster, information is shared about who needs help, the locations of vulnerable people, and the regions with substantial damage (Dragović et al., 2019). For example, volunteers from Tufts University after the earthquake in Haiti created a map that has helped survivors and volunteers sent rescue information via messages on Twitter. Within 15 days, more than 2.500 messages were received (Fraustino J.D. et al., 2012; Gao H. et al., 2011).

Natural disasters	year	City/region/country	Social media	Authors
Earthquake	2010	Haiti, Caribbean	Twitter	Lobb et al., 2012
Hurricane	2011	Fairfax County, Virginia	Twitter, Facebook, YouTube	Fraustino et al., 2012; Slide Share, 2012
Flood and Landslide	2014	Kashmir, Indonesia	Twitter	Chaturvedi et al., 2015

Table 3: Case studies from recovery phase (Dragović et al., 2019).

The use of tweets as an indication of the spatial footprint of a phenomenon is also possible. For example, Acar and Muraki, (2011) examine the use of Twitter during an earthquake in Japan. They observed that tweets from affected areas include requests for help and warnings. Tweets from other areas far from the disaster epicentre tended to include other types of information, such as messages of concern and condolences. Moreover, messages are highly heterogeneous, with multiple sources (e.g., institutional accounts, media, eyewitness accounts, influencers, and bots) and varying levels of quality (Imran et al., 2015). Additionally, different languages can be utilized in the same crisis, particularly if there are events impacting several geographic contexts, such as transborder areas.

Most disastrous event detection systems are confined to detecting whether a tweet is related to a disaster based on textual content (Singh et al., 2019). Event detection is usually performed by discovering unusual activity patterns focused on a particular geographic area or on a given topic (usually specified by means of keywords). Recently, there has been growing interest in machine learning natural language processing (NLP). Several research communities realize many labelled datasets for different events and tasks, such as text analysis about sentiment analysis (SA), opinion mining or topic modelling. The following resources are made available to help researchers and technologists advance research on humanitarian and crisis computing by developing new computational models and innovative techniques. Therefore, analysing the content of micro-posts may be useful in the selection of only those that are relevant to a determined task. In the case of disaster management, identifying posts that indicate a situation of danger, worry or generic alarm may suggest important (Buscaldi et al., 2015).

2 MATERIAL

This chapter provides a detailed description of the different data sources for mapping Italy. The datasets cover 10 years, from 2010 to 2019, for the entire Italian territory. Each database identifies one or more effects as a consequence of an event from published articles in real-time, data rainfall, human lives and earmarked funds. Greater importance was placed on the collection of landslide event information considering three hazard maps (ISPRA).

Finally, a step forward, with a data mining technique within Twitter, is described. The data from Twitter combined with the deep learning of information for text classification represent further exploitable data that is available during natural disasters. Data mining, natural language processing, deep learning and statistical analyses have been carried out using Python programming, to which a brief part of this paper has been dedicated.

2.1 Study area

Italy is almost 300.000 km^2, and it is divided into 20 regions (Figure 1A), with 107 provinces and 7926 municipalities. Furthermore, 158 Warning hydrological zones (WHZs) have been outlined on the basis of morphology, catchment boundaries and administrative limits. Much of Italy consists of hilly and mountainous terrain subject to landslides of different types and sizes (Guzzetti, 2000). The main mountain chains are the Alps in North Italy and the Apennines (Figure 1B), which span from north to south. In the alpine area, which is formed mainly of metamorphic rocks (Vai & Martini, 2001, Salvatici et al., 2018), the most frequent phenomena are rock falls and debris avalanches (Agliardi & Crosta, 2003; Panizza et al.,2011), while in the Apennines, which are formed mainly of arenaceous flysch (Vai & Martini, 2001; Agostini et al., 2014; Rosi et al., 2018, 2020), the most common landslides are represented by rotational and translational landslides, both surficial and deep-seated. The climate of Italy is mainly Mediterranean, with dry and warm summers and mild and wet winters; during winter, snowfall is frequent both on the Alps and on the Apennines, and the consequent snowmelt in the spring often leads to the mobilization of landslides.

Italy is the European country with the widest area distribution and the highest recurrence of large landslides, causing severe losses of lives and goods (Salvati et al., 2010; Avvisati et al., 2019). Currently, the IFFI database (Italian Inventory of Landslides; Trigila et al., 2007) includes more than 600.000 landslides affecting an area of 23.700 km^2, representing 7,9% of the national territory (Trigila and Ladanza, 2018). Every year, thousands of landslides occur in the national territory, and a few hundred of these create victims, casualties, evacuations and damage to buildings, cultural heritage, and the primary transportation infrastructure. For example, in 2017 172 events were reported and there were 146 in the previous year (Trigila and Ladanza, 2018). Legambiente, (2021) surveyed 1181 extreme weather events from 2010 to the present that caused damage in Italy. A total of 637 municipalities (8% of the total) recorded events with relevant impacts. In terms of human lives and injuries, 264 people have been victims of natural disasters. The CNR (National Research Council) recorded the evacuation of over 27.000 people due to events such as landslides and floods between 2016 and 2020, which becomes 320.000 when counting the events that have occurred since 1971. The regions most affected by extreme events since 2010 are Sicilia and Lombardia, with 144 and 124 events, respectively (Franceschini et al., 2022a and b).

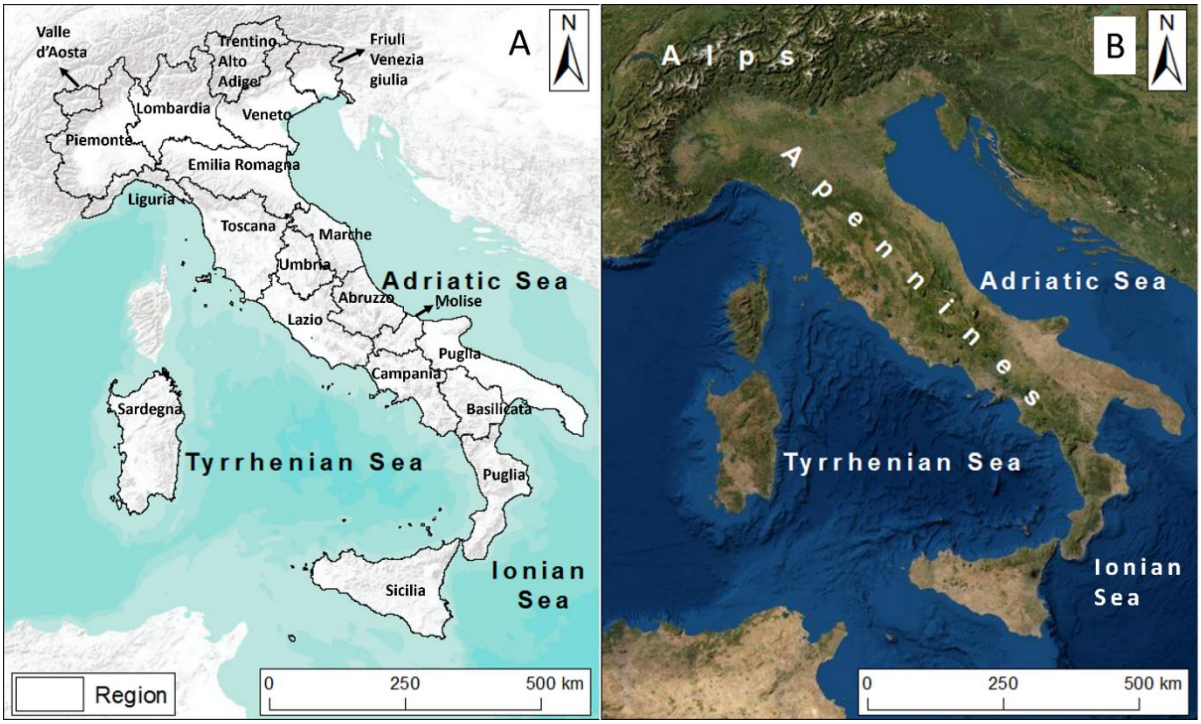

Figure 1: A Regions of Italy. **B** Image from the satellite of Italy and its main mountain chain. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

2.2 Preliminary data sources

Different sources of information have been analysed in this work. Several organizations create reports or datasets describing many different aspects of natural disasters. In the beginning, newspaper articles were analysed and correlated with other available source data.

In this work, four 10-year-long datasets (2010-2019) of landslide events and flood events in Italy were analysed. The analysis was carried out to obtain information and to determine the spatial, regional and temporal correlations of the available data. Furthermore, additional informational maps from ISPRA have been utilized for landslide events.

Below, several datasets are used to obtain information about landslide and flood events:

1. Newspapers can be used to create a landslide inventory, which, in turn, can be analysed for landslide hazard assessments (e.g., landslide distribution, frequency and intensity).
2. Rainfall data allow us to obtain the number and frequency of rainfall events;
3. Populations at risk from landslides and floods in Italy (Polaris) identifies the event effects in terms of human lives and involved regions;
4. The National Repository of Soil Defence interventions (ReNDiS) inventory outlines earmarked funds for soil protection;
5. Maps of the percentage of landslide hazard areas, percentage of people at risk and percentage of buildings at risk (ISPRA);

Polaris and ReNDiS identify the event effects in terms of human lives, involved regions and earmarked funds for remediation and risk mitigation works.

For flood events, only the first four points were utilized. For landslide events, a more detailed focus has been applied using all the oversized points also to check and/or validate the data.

Some parts of the statistical approach were carried out using MATLAB R2021b and Python. ArcMap and ArcGISPro provided by ESRI have been utilized to create several maps.

Newspaper articles dataset

The data research took place within one main news aggregation platform (Google News), harvesting 184.322 articles about landslide events (Figure 2A) and 246.338 about flood events (Figure 2B) from 2010 to 2019. The retrieved articles have been grouped based on the event they refer to. In this way, 32.525 landslide and 34.560 flood event news items were identified. The news database was classified into three classes on the basis of news relevance, localization accuracy and time of publication.

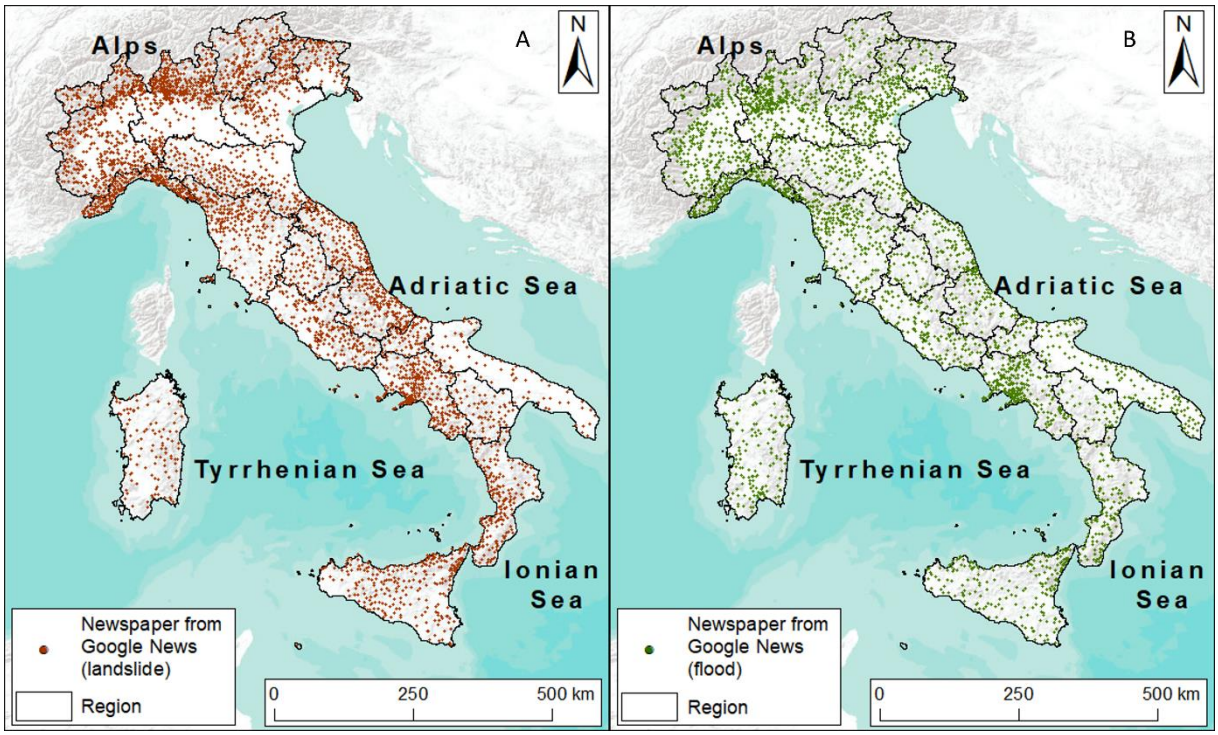

Figure 2: General distribution of the news used in Italy for Landslide (**A**) and Flood events (**B**). The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

Rainfall dataset

A dense network of spatially distributed rain gauges over Italy provides continuous direct observations of the rain measurements for several specific locations (Colle et al., 1999). This network is made up of over 4500 rain stations, which provide rainfall measurements approximately every 15 minutes (Figure 3A). This network was deeply analysed by Del Soldato et al., (2021). Each pluviometer was analysed to select only the rain gauges recording data for more than 20 h per day (to remove the data with low representativeness) and to remove noisy data (e.g., negative rainfall values or higher than 400 mm/h). In this way, a robust database from a statistical point of view was set. The analysis was carried out from data covering the period 2010-2019. The same authors divided the rainfall events into five classes (in accordance with the classification used by the Italian Civil Protection Department) based on their average intensity (as mm/day). For each class, the number of events over the analysed period was calculated (Figure 3B).

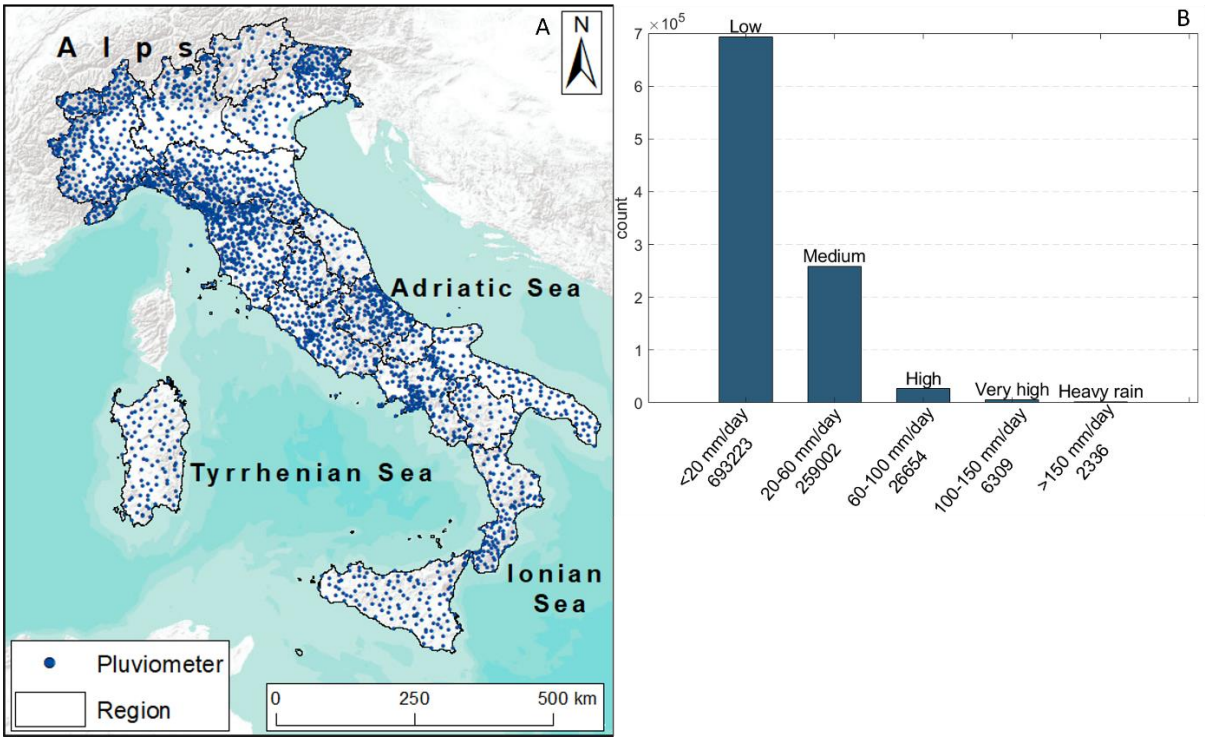

Figure 3: **A** Rain gauges across the Italian territory and in **B** Rainfall data between 2010 and 2019 (from Del Soldato et al., 2021). For each class, the number of occurrences was calculated (event count). The map was generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8). The panel was generated using MATLAB R2021b.

Populations at risk from landslides and floods in Italy Polaris

Polaris (Popolazione a Rischio da Frana e da Inondazione in Italia - Populations at risk from landslides and floods in Italy) is a website managed by the Research Institute for Hydrogeological Protection (IRPI) of the National Research Council (CNR) of Perugia (Italy). In the attempt to assess the geo-hydrological risk to the Italian population, for years, IRPI collected and processed historical information on landslides and floods that have caused damage to the population. Every year, Polaris produces a report with various statistics on the distribution of the fatalities and casualties due to these natural events. For each considered event (landslide and flood), the database provides information about the involvement of municipalities, provinces, regions, victims, casualties and people who are missing and evacuated. The dataset has information from 2011, the year in which the project started, and no information was collected for 2010.

The National Repository of Soil Defence interventions-ReNDiS

ReNDiS is a database of remediation works planned to repair the damages derived from natural events such as landslides and floods; it was founded by the Italian government (Campobasso et al., 2013).

Data are collected through continuous contact between ISPRA (*Istituto Superiore per la Protezione e la Ricerca Ambientale* - Italian Institute for Environmental Protection and Research) technicians and the local authorities managing the works in the Italian territory. Therefore, through the ReNDiS database, the Italian government can be informed in real-time of how their funds for risk mitigation work are being spent and how they are distributed across the country (Campobasso et al., 2013). The ReNDiS dataset starts in 2000 and covers the period until 2020. The years refer to the financed interventions. Overall, more than 2,7 billion euros were allocated as funding to remediate the structural damage and economic losses. For the analysis conducted in this work, the year of intervention funding, from 2010 to 2019, the involved region, the landslide event (Figure 4A) and flood event (Figure 4B), and the incurred expenses were considered.

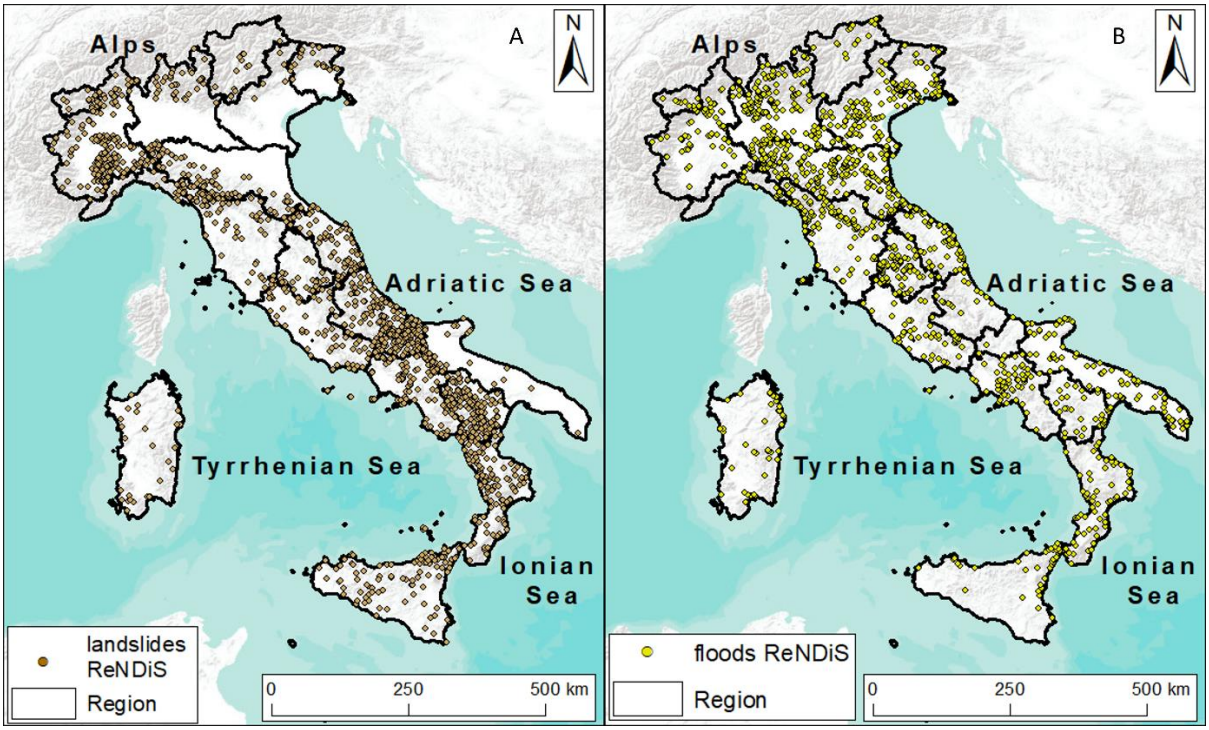

Figure 4: **A** ReNDiS data for landslide events and **B** ReNDiS data for flood events. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

Hazard maps

Hydrogeological instability is an issue of particular relevance for Italy because of the impacts on the population, environment and infrastructure. Italy is a strongly anthropized country with a natural propensity to instability, linked to its climatic, topographic, morphological and geological characteristics. A landslide hazard represents the probability of occurrence of a potentially destructive phenomenon of a given intensity in a given period and a given area (Varnes, 1984). The landslide hazard areas of Piano Assetto Idrogeologico (PAI) include, in addition to the landslides that have

already occurred, the areas of possible evolution of the phenomena and the areas potentially susceptible to new landslide phenomena (Trigila et al., 2021). ISPRA instituted new maps (2020-2021), collecting several analyses of landslide hazards from different institutions (regions, autonomous provinces and catchment authorities). ISPRA uses 5 classes to classify (very high hazard P4, high hazard P3, medium hazard P2, moderate hazard P1 and attention areas AA) the landslide hazard and creates PAI maps for the entire national territory. In this work, the sum of the percentages of P3 and P4 was considered (Figure 5A). In particular, we highlight the importance of high landslide hazards in some areas. The hazard landslide area P4 is approximately 9595 km^2 (3,1%), and that in P3 is almost 16.891 km^2 (5,6%).

The population at risk is defined as the population living in landslide hazard areas exposed to the risk of personal injury (dead, missing, injured, evacuated). A total of 500.000 people live in very heavy hazard areas (P4), and almost 804.000 live in high-hazard areas (P3) (Figure 5B). Hence, 1,3 million people live in areas with high hazard levels, approximately 2,2% of the total (59 million).

The buildings at risk in P3 and P4 are 565.000 and are almost 3,9% of the total (12 million buildings in Italy) (Figure 5C).

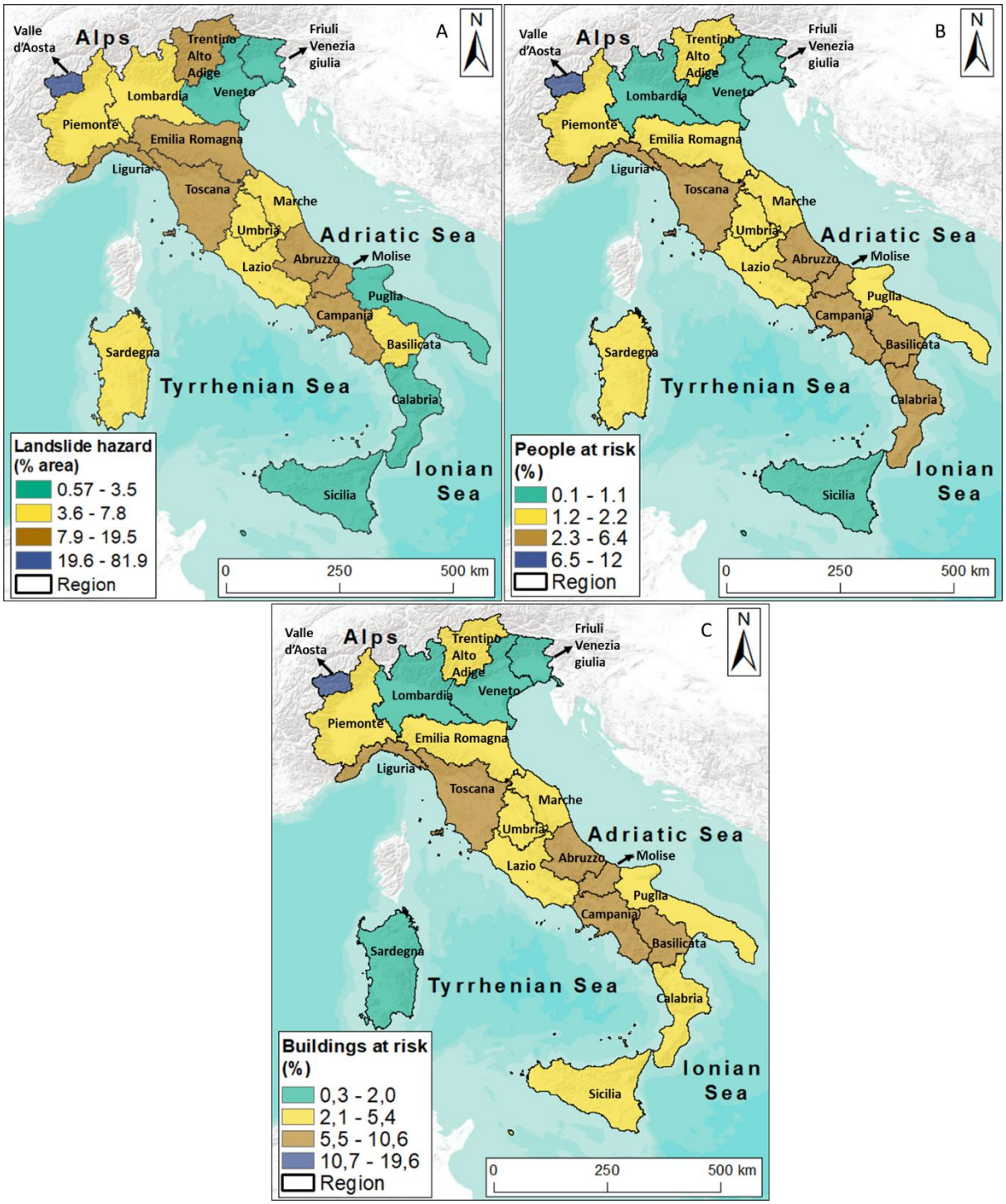

Figure 5: Piano Assetto Idrogeologico (PAI) landslide hazard areas on a regional basis classified areas into four labels (**A**). Percentage of people at risk on a regional basis classified areas into four labels in **B**; Percentage of buildings at risk on a regional basis classified areas into four labels in **C**. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

2.3 Data from Twitter

This section includes a step forward with the data mining technique within Twitter.
Twitter now has almost 400 million active monthly users, meaning a huge volume of data is available to collect, most of which is public.
The data mining technique has been applied only to landslide events. The results from a newspaper landslide analysis allowed us to obtain the main keywords. The keywords have been utilized to get tweets from Twitter during specific periods. The 5 keywords utilized for data mining are shown in Table 4.

Keywords	frana OR smottamento OR scivolamento OR crollo OR dissesto

Table 4: Requests used to collect tweets from the Twitter academic API2.

The Tweets were collected through the Twitter academic API2 with an academic licence. Overall, 13.349 tweets were harvested. The periods have been considered based on the temporal distribution of news. The dataset obtained can represent:

- a landslide inventory from which to obtain information about some landslide events;
- a point of comparison between news and tweets;
- a solid base to apply deep learning analysis.

Table 5 shows slots for 10 years and the relative extracted tweet number.

year	from	to	data
2011	22/03/2011	22/03/2011	1
	01/10/2011	30/11/2011	420
2012	01/09/2012	31/12/2012	693
2013	01/01/2013	31/05/2013	1028
2014	01/01/2014	31/05/2014	1747
	01/07/2014	30/11/2014	1319
2015	22/02/2015	26/02/2015	1626
2016	24/11/2016	28/11/2016	1656
2017	05/08/2017	08/08/2017	486
2018	28/10/2018	31/10/2018	2273
2019	24/11/2019	25/11/2019	2100

Table 5: Many slots over 10 years have been extracted using five words. The periods were chosen on the basis of the temporal distribution of newspaper articles.

From Twitter it is possible to obtain different metadata such as User object, Tweet Object and Place objects. The User object contains Twitter user account metadata describing the referenced user. Below many fields are listed and considered significant during data mining:

- fid: Field identity. The format is a string.
- author_id: The unique identifier of the User who posted this Tweet. The format is a string ("author_id": "2244994945").
- name: The name of the user, as they've defined it on their profile. Not necessarily a person's name. Typically capped at 50 characters, but subject to change. The format is a string ("name": "Twitter Dev").
- username: The Twitter screen name, handle, or alias that this user identifies themselves with. Usernames are unique but subject to change. Typically a maximum of 15 characters long, but some historical accounts may exist with longer names. The format is a string ("username": "TwitterDev").
- author_created_at: The UTC datetime that the user account was created on Twitter. Can be used to determine how long a someone has been using Twitter. The format is a date (ISO 8601) ("created_at": "2013-12-14T04:35:55.000Z").
- author_description: The text of this user's profile description (also known as bio), if the user provided one. The format is a string ("description": "The voice of Twitter's #DevRel team, and your official source for updates, news, & events about Twitter's API. \n\n#BlackLivesMatter").

- author_entities: Contains details about text that has a special meaning in the user's description. Entities are JSON objects that provide additional information about hashtags, urls, user mentions, and hashtags associated with the description. Reference each respective entity for further details. All users start indices are inclusive, while all user end indices are exclusive. Contains details about text that has a special meaning in the user's description. The format is object.
- author_location: The location specified in the user's profile, if the user provided one. As this is a freeform value, it may not indicate a valid location, but it may be fuzzily evaluated when performing searches with location queries. The format is a string ("location": "127.0.0.1")
- public metrics: author_followers, author_following, author_tweet_count and author_listed_count. Contains details about activity for this user. Can potentially be used to determine a Twitter user's reach or influence, quantify the user's range of interests, and the user's level of engagement on Twitter. The format is object ("public_metrics": { "followers_count": 507902, "following_count": 1863, "tweet_count": 3561, "listed_count": 1550 }).
- author_url: The URL specified in the user's profile, if present. A URL provided by a Twitter user in their profile. This could be a homepage, but is not always the case. The format is a string ("url": "https://t.co/3ZX3TNiZCY")
- author_verified: Indicates if this user is a verified Twitter User. Indicates whether or not this Twitter user has a verified account. A verified account lets people know that an account of public interest is authentic. The format is boolean ("verified": true).

The Tweet object has a long list of 'root-level' fields, such as id, text, and created_at. Below only some parameters have been listed and in consequence considered for extracting and relevance for describing tweet text:

- id_text: The unique identifier of the requested Tweet. Use this to programmatically retrieve a specific Tweet. The format is a string ("id": "1050118621198921728").
- text: The actual UTF-8 text of the Tweet. Keyword extraction and sentiment analysis/classification. The format is a string.
- created_at: Creation time of the Tweet. This field can be used to understand when a Tweet was created and used for time-series analysis etc. The format is date (ISO 8601) ("created_at": "2019-06-04T23:12:08.000Z").
- lang: Language of the Tweet, if detected by Twitter. Returned as a BCP47 language tag. Classify Tweets by spoken language. The format is a string ("lang": "en").
- entities: Entities that have been parsed out of the text of the Tweet. Entities are JSON objects that provide additional information about hashtags, urls, user mentions, and hashtags associated with a Tweet. Reference each respective entity for further details. All start indices are inclusive. The

majority of end indices are exclusive, except for entities.annotations.end, which is currently inclusive. The format is object.

- public_metrics: retweets, replies, likes and quote_count. Public engagement metrics for the Tweet at the time of the request. Use this to measure Tweet engagement. The format is an object ("public_metrics":{ "retweet_count": 8, "reply_count":2, "like_count":39, "quote_count": 1 })
- in_reply_to_user_id: If the represented Tweet is a reply, this field will contain the original Tweet's author ID. This will not necessarily always be the user directly mentioned in the Tweet. Use this to determine if this Tweet was in reply to another Tweet. The format is a string ("in_reply_to_user_id": "2244994945").
- source: The name of the app the user Tweeted from. Determine if a Twitter user posted from the web, mobile device, or other app. The format is a string ("source": "Twitter Web App").

The Place objetcs tagged in a Tweet are not a primary object on any endpoint, but can be found and expanded in the Tweet resource. In this sense, only field have been examined:

- Geo: Contains place details in GeoJSON format. The format is an object ("geo": { "type": "Feature", "bbox":[-74.026675, 40.683935, -73.910408, 40.877483], "properties": {}}).

3 METHODS

Different methods have been applied to obtain information about natural events from different data sources. The workflow can be split into three parts: i) state-of-the-art analysis; ii) upgrade, processing and analysis data from Twitter; and iii) results made up of spatial information and temporal distributions (Figure 6).

In the first part of the state-of-the-art, newspaper articles from Google News were considered the starting point for several analyses. A dataset of newspaper articles was classified manually and then used as a proxy for landslide and flood hazard estimation. Three other datasets were used to obtain information about events: rainfall data, affected people and the reported expenses on the soil protection measures. The analysis allowed us to obtain spatial and temporal information about the natural events (Figure 6 in blue) and their effects in terms of human losses and the earmarked funds for the entire Italian territory. Subsequently, in the second step using the headlines, a Natural Language Processing (NLP) technique (Liddy, 2001) was applied to obtain the word frequency. The word frequency technique has been used with the intention of identifying the most common associations of words for all the news.

Given the present literature on data mining for flood events and the absence of studies on landslide events, the following passages focus on the latter events to deepen and analyse a topic not truly addressed in social media analyses and crowdsourcing platforms. The first five words most frequently identified were keywords. The five keywords from the headlines have been used to apply the data mining technique within Twitter. The period was chosen on the basis of temporal analysis in the first step. The achieved dataset has been manually classified to create a solid basis for the deep learning techniques. The transformer architecture has been chosen for text classification within deep learning (Figure 6 in orange). The method that has been chosen is the XLM-RoBERTa with model "xlm-roberta-base".

Finally, the results are shown throughout panels and maps, outlining the spatial and temporal distribution of natural events from newspaper articles, their effects, and their correlation with hazard maps and Twitter for the entire Italian territory for 10 years (Figure 6 in green).

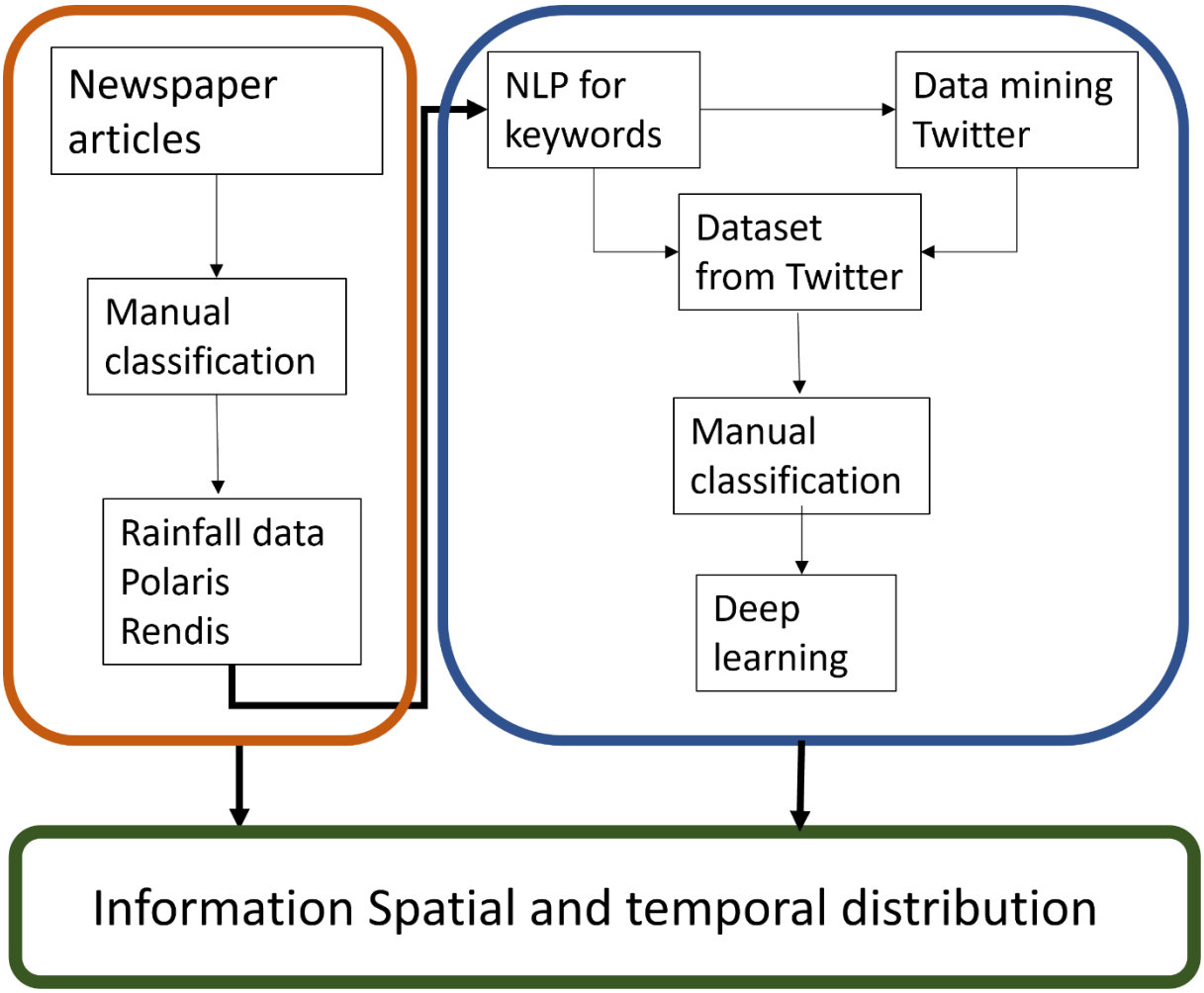

Figure 6: The workflow split in three ways: in orange, the first path with several analyses about state of art within the paper article; in blue, the second path with an upgrade for the data mining inside the crowdsourcing platform during a natural disaster in a particular landslide event. Both paths were analysed to obtain information about the spatial and temporal distribution of the entire Italian territory for 10 years (2010-2019).

3.1 Newspaper articles

The Semantic Engine to Classify and Geotagging News-SECaGN is an algorithm based on a mechanism of acquisition, management and publishing of online articles related to natural hazards (landslides, floods and earthquakes). It aims to obtain information about the spatial and temporal distribution of the events. An automatic search for newspaper articles is performed combining primary words, synonyms, singular and plural forms (keywords) in the Italian language related to the landslide argument. Data mining is applied inside Google News. After the acquisition process, a data filtering procedure is applied to separate the nonrelevant information from the pertinent items. Data filtering takes place through geotagging and cataloguing the articles using three scores (Battistini et al., 2013):

- Place score: A score value is assigned to evaluate the reliability of the geotag;
- Event score: Index of the probability that the news item actually concerns the topic event;
- Time score: Estimated days between the time of occurrence of the event and the time of publication of the article.

All the newspaper articles that reach a minimum score are then filed in a geodatabase, and their location can be viewed in a dedicated WebGIS. The whole process is repeated every 15 minutes.
This data mining methodology was calibrated and tested in Italy during a test period of 2 years (November 2009 – November 2011). The process is completely automated and scalable. It can also be applied in other countries after a specific tuning of the keywords used by the data-mining algorithm.
In this work, the news database underwent manual classification based on news relevance, localization accuracy and time of publication. This classification allows us to identify the most relevant news in terms of the temporal and spatial accuracy of the landslide event identification.
For the classification, 3 classes have been defined (Table 6):

Class 1: "Near real-time news". In this category, all the news referring to the ongoing or very recent landslide and flood events (same day or a couple of days before) are classified. This news is also characterized by a high level of spatial accuracy (at least the municipality must be identified), with an approximation of a few kilometres. Some news, with high temporal precision but low spatial accuracy, have been manually modified (if possible) based on article text to reach the required level of approximation. The news in this class can be used for further analyses or modelling (Battistini et al., 2017).

Class 2: "News generically referring to an event". In this category, the news referring to past landslide or flood events with unknown triggering dates (e.g., the initiation (or finishing) of works aimed at risk reduction or landslide remediation) is stored. News with low spatial accuracy (referring to provinces/cities or geographical areas) is classified in Class 2 as well. This kind of news is useful for identifying those areas that have been affected by landslides or floods in the past, and for risk zoning.

"News not related to event". News not related to the landslide or flood argument but whose semantic association leads to a misclassification. After this work, these news items were removed from the database.

Class	Time	Localization	Title example
1	Near real-time	Municipality or village	"Gallivaggio landslide, the video" "Piemonte, floods in Alessandria Province"
2	The date of the event cannot be defined	City, region, lake, river	"The funds for securing the landslides are now available, the works will start soon" "Flood emergency, work to reopen the Bulagaio."
3	-	-	Italy's economy doesn't grow up, the South slide down

Table 6: Description of the 3 classes used to group the news.

3.2 Newspaper articles and traditional sensors

Four datasets were analysed and compared to each other to assess the distribution and evolution of the landslide and flood hazards and their effects in the Italian territory (Figure 7). Each dataset was analysed, filtered and homogenized to obtain possible correlations. The output has been made up of panels and maps, which describe the temporal and spatial distribution of landslide events in Italy over 10 years.

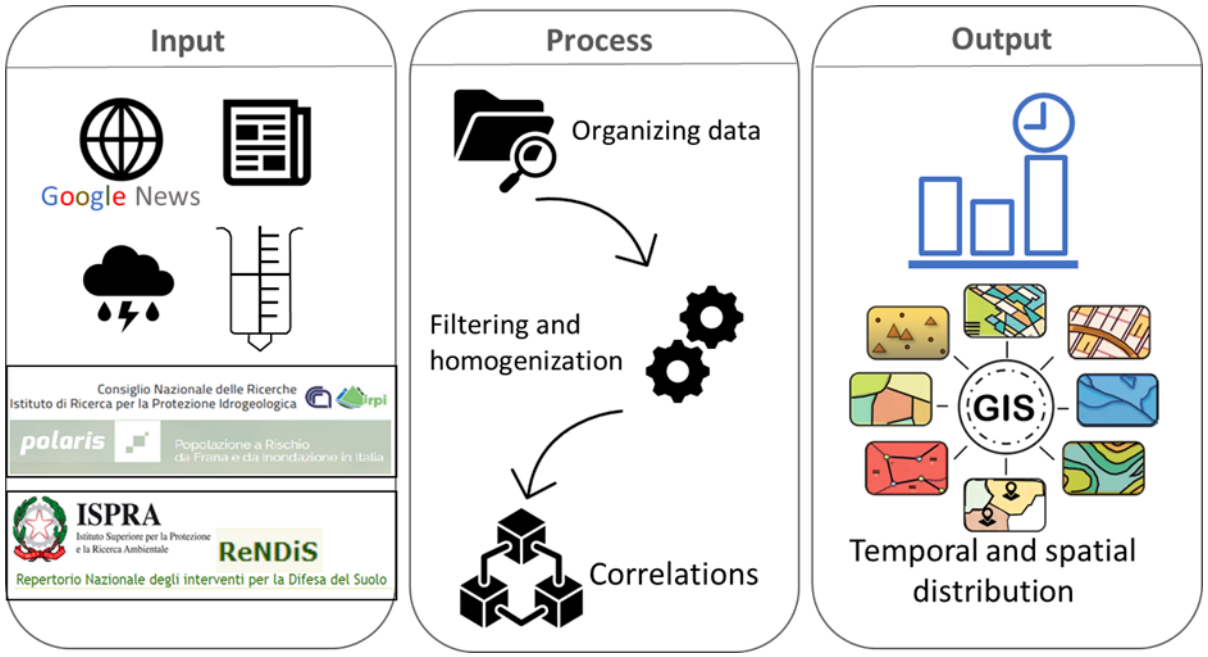

Figure 7: Workflow of the work. The output corresponds to panels and maps to obtain information on approximately 10 years of landslides for the whole Italian territory.

Each dataset provides specific kinds of data, which are organized as described below:

- "Landslide news" and "Flood news" - The social media database contains two pieces of information: (i) "Landslide news", "Flood news" and (ii) "Newspaper articles" for each event. Landslide news and flood news refer to articles considering the same event and grouped into a single data item. This information can outline the hazard of a certain area; thus, the higher the number of "Landslide news" or "Flood news" items, the higher the propensity to hazard an area. "Newspaper articles" is the sum of several articles published for each event from different newspapers. The number of contributions outlines the media impact of the event; the higher the number of publications, the higher the intensity and the impact of the event.

The summarized targets and relative nomenclature are in Table 7. Each target identified a meanly and in consequence parallel considerations.

Target	Identified	Parallel results
Class 1	Landslide event Flood event	Landslide day Flood day
Class 1+2	Landslide news Flood news	Estimate Landslide hazard Estimate Flood hazard
Total articles	Newspaper articles	Media impact

Table 7: Nomenclature of each target.

- Rainfall data - they have been classified by Del Soldato et al., 2021 into five classes based on daily intensity. Among these classes, "High intensity" (60-100 mm/day), "Very high intensity" (100-150 mm/day) and "Heavy rain" (> 150 mm/day) show a clearer spatial distribution with respect to "Medium intensity" (20-60 mm/day) and "Low intensity" (5-20 mm/day). For this reason, the frequencies of the rainfall events of these classes have been considered (named "relevant rainfalls" hereafter), obtaining two databases, the frequency of each class of intensity and the count of the rainfall events.

- Polaris - two ranks of affected people were distinguishing human involvement for the spatial and temporal distributions: (i) Injured, Deaths, Evacuated and Missing people (IDEMs) used for the temporal distribution analyses of human involvement and (ii) Injured, Deaths and Missing personnel used for the spatial distribution investigation. The difference between these two groups is due to the lack of spatial information for the evacuated people.

The first step of the analysis aims to outline the temporal evolution of each dataset from 2010 to 2019. Then, each dataset was analysed to obtain its spatial distribution at the regional and WHZ (Warning Hazard Zone) scales.

Regional scale analysis attempts to provide an overview of the spatial distribution of the considered variables: (i) ”Landslide news” and “Flood news”, (ii) “Newspaper articles” for each event, (iii) rainfall intensity, (iv) rainfall events, (v) IDMs, and (vi) funds. Then, the percentage of hazardous landslide areas and the percentage of buildings at risk (ISPRA, 2021) in each region were correlated with the earmarked funds for soil protection. The percentages have been scaled on the basis of the regional size with respect to the Italian territory (300.000 km^2) and to the building numbers for the entire national territory (12.187.698 buildings according to the National Institute of Statistics) to proportionate the results with the size and urbanisation of the region.

At the WHZ scale, a more detailed analysis of the landslide news and the rainfall events has been made. Regarding rainfall data, the frequency of the 3 intensity classes, “high”, “very high” and “heavy rain”, and the number of events of each class (event count) were correlated.

3.2.1 Keyword extractions inside headline newspaper articles

A preliminary processing of the data for the semantic analysis was conducted to check and assess, inside the headlines, the frequency of words that describe the landslide events. In this way, the most common association of words both for “good” and “bad” news is identified and can be used to improve the system. The headlines of each article have been analysed using the Natural Language Processing (NLP) technique (Liddy, 2001). NLP is a computerized approach to textual analysis, and it provides several techniques to model textual data. In this work, the word frequency technique has been used with the intention of identifying the most common associations of words both for “good” and “bad” news. The results of this analysis can help to improve the data mining algorithm.

3.3 Data mining within Twitter

Crisis situations, such as natural disasters, generate a situation that is rife with questions, uncertainties, and the need to make quick decisions, often with minimal information (Imran et al., 2015). In regard to information scarcity, research in recent years has uncovered the increasingly important role of social media communications in disaster situations and has shown that information broadcast via social media can enhance situational awareness during a crisis situation (Vieweg, 2012). There is a

recognition that social media communications are a valid and useful source of information throughout the disaster lifecycle (preparation, impact, response and recovery) (Imran et al., 2015).

The use of social media to communicate timely information has become a common practice in recent years. In particular, the one-to-many nature of Twitter has created an opportunity for stakeholders to disseminate crisis-relevant messages (Olteanu et al., 2015).

In general, crowdsourcing platforms provide application programming interfaces or APIs. An Application Programming Interface is a software intermediary that allows two applications to communicate with each other to access data (Figure 8). Essentially, developers plug into APIs to access certain assets for the end users. Twitter provides two Application Programming Interfaces (APIs): i) Search APIs allow us to obtain queries of an archive of past messages; ii) Streaming or filtering APIs allow data collectors to subscribe to a real-time data feed. Both types of APIs typically allow data collectors to express an information need, that includes one or several of the following constraints: (i) a time period; (ii) a geographical region for messages that have GPS coordinates (which are currently the minority); or (iii) a set of keywords that must be present in the messages, which requires the use of a query language whose expressiveness varies across platforms. In the case of archive/search APIs, messages are returned sorted by relevance (a combination of several factors, including recency) or just by recency. In the case of real-time/streaming/filtering APIs, the messages are returned in the order of their posting time (Imran et al., 2015). In general, the Twitter API enables programmers to access Twitter in advanced ways. It can be used to analyse and interact with tweets.

In the second half of 2020, the Twitter developer team rebuilt the Twitter API, releasing Twitter API v2.

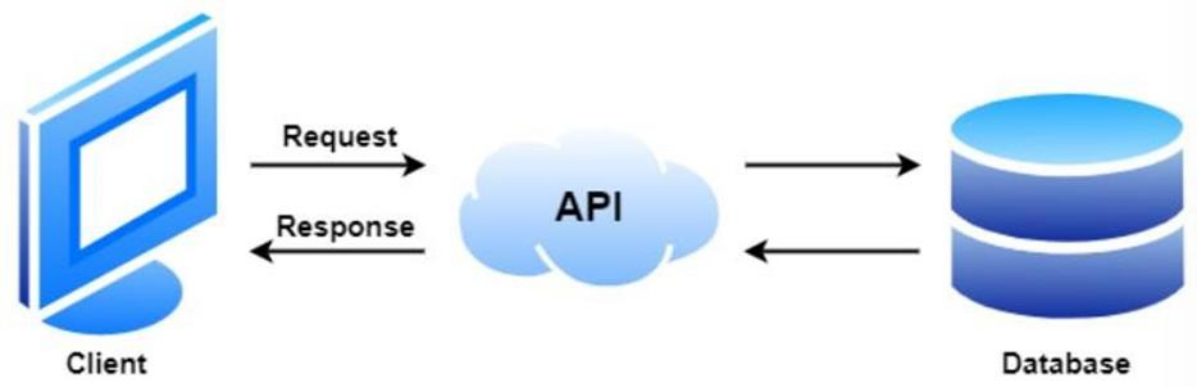

Figure 8: A web-based API that takes in a client's request and returns data in response.

In this work, tweets were collected through the Twitter academic API2 with an academic license. On the developer portal of Twitter, consumer keys with the *API Key and Secret* and authentication tokens with *Bearer Token* and *Access Token and Secret* were acquired for access. At the same time, an academic project has been created with the name "Data mining during natural disasters". Such a

project is supported by an application named “Data mining for landslides”. Academic access allows us to obtain 10 million tweets per month.

Data mining was carried out using Python programming. Python is a programming language. It has syntax rules for writing one code, which will be considered valid by Python interpretation software that will read and run instructions. It allows one to work quickly and integrate systems more effectively. This work environment has a wide range of syntactical constructions, standard library functions and interactive development environment features. Currently, Python is the most widely utilized language in scientific computing.

3.4 Machine learning for text analysis

Machine learning is a method of data analysis; it is part of artificial intelligence (AI) that is based on the idea that systems can learn from data, outline correlations and make decisions with minimal human involvement.

Three widely adopted machine learning methods are supervised learning, unsupervised learning and semi-supervised learning.

Supervised learning (SL)

Supervised learning, during training, uses labelled data to learn. These datasets are designed to train algorithms into classifying data or accurately predicting outcomes accurately. Using labelled inputs and outputs the model can measure its accuracy and learn over time.

SL can be split into two different models:

- Regression uses an algorithm to understand the relationship between dependent and independent variables. Regression is most often used to predict numerical values based on previous data observations. Popular regression algorithms include linear regression, logistic regression and polynomial regression;
- Classification uses an algorithm to accurately assign test data into specific categories, such as separating dogs from cats. Common types of classification algorithms are linear classifiers, support vector machines, decision trees and random forests.

Unsupervised learning (UL)

Unsupervised learning uses machine learning algorithms to analyse and cluster unlabelled datasets. These algorithms extract hidden patterns in data without the need for human assistance.
UL can be split into three different models:

- Clustering is a data mining technique for grouping unlabelled data based on their similarities or differences. For example, K-means clustering algorithms group similar data points, where the K value is the size of the grouping. This technique is helpful for market segmentation, image compression, etc. (International Business Machines Corporation-IBM);
- Association uses different rules to determine the relationship between two variables in a dataset;
- Dimensionality reduction is applied when the number of features in a dataset is too high. Reducing the number of data inputs allows better management and preservation of the data integrity. This technique is employed in the data preprocessing stage, as autoencoders remove noise from visual data to improve picture quality (IBM).

Semisupervised learning

Semisupervised learning falls between SL and UL; it involves a small portion of labelled examples and many unlabelled examples from which a model must learn and make predictions on new examples. This type of learning can be utilized for classification, regression and prediction.

Integrated into the theme of artificial intelligence between machine learning and deep learning are natural language processing and text analytics. Both use machine learning algorithms to understand the meaning of text documents and speech. The role of natural language is to improve, accelerate and automate the functions, modifying unstructured text into useable data.

3.4.1 Natural Language Processing

Natural language processing is an interdisciplinary area within artificial intelligence among machine learning-deep learning, text analysis and computational linguistics (Figure 9).

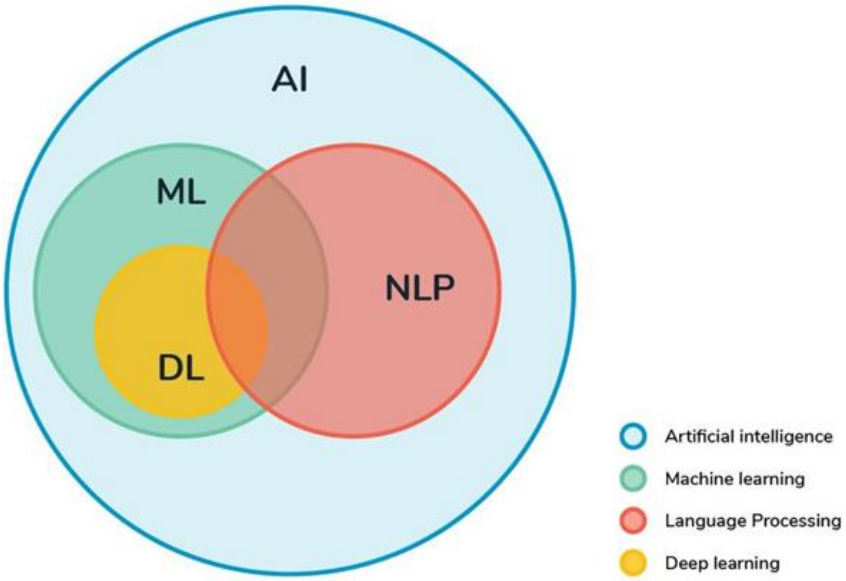

Figure 9: Interdisciplinary natural language processing activity within artificial intelligence, involving techniques of machine learning and deep learning.

Natural language processing helps machines read, understand, replicate and derive meaning from human languages. On the other hand, NLP is aimed at understanding the linguistic use and context behind the text, analysing grammatical structures and semantics (Figure 10). In general, natural language processing is utilized to analyse large volumes of text data, such as social media, comments, reviews, and news reports. Despite the success of neural models for NLP tasks, the performance improvement may be less significant compared with the computer vision (CV) field (Qiu et al., 2020). NLP is an integral part of technology such as Google Translate, voice assistants (Alexa, Siri, etc.), chatbots, Google searches, and voice-operated GPS.

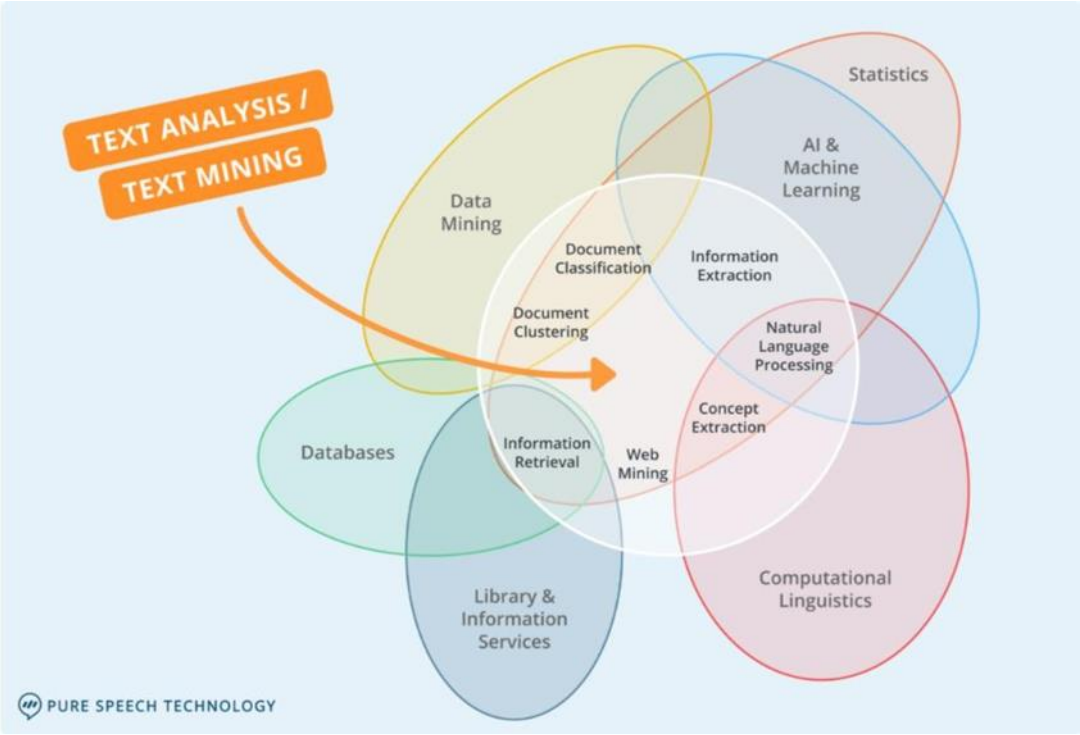

Figure 10: Venn diagram showing the intersection of text analysis (or text mining) with six related fields: statistics, AI and machine learning, computational linguistics, library and information services, databases and data mining Miner et al., (2012).

A typical workflow of text analysis and thus of the application of natural language processing is depicted in Figure 11. The collected data can derive from various sources, as long as they are textual data. Data can be internal if they derive from emails, chats, and surveys, while external data are obtained from social media, news, and online reviews.

Preprocessing represents the first processing step, which allows qualitative and quantitative information to be obtained from textual data. Preprocessing allows unstructured (textual) data to be transformed into ordered data, i.e., into a sequence of numbers.

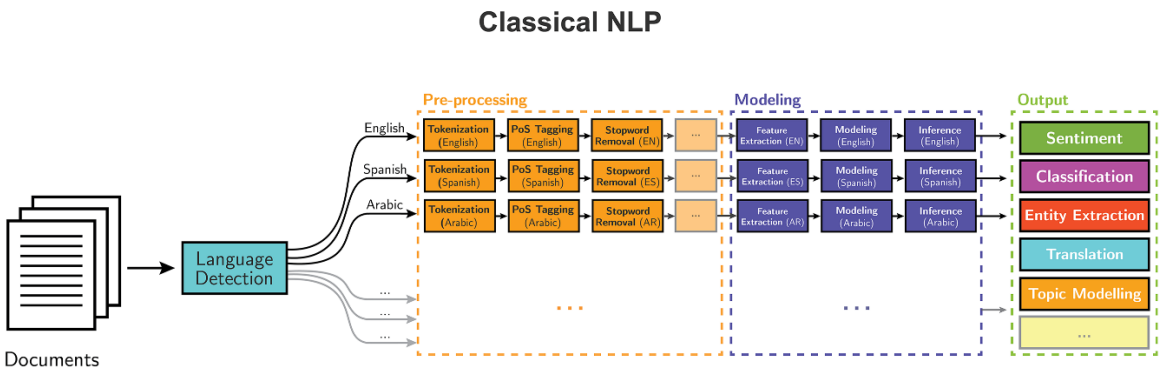

Figure 11: Classical NLP approach (Image credit: https://s3.amazonaws.com/aylien-main/misc/blog/images/nlp-language-dependence-small.png).

Subsequently, text analysis is applied as a modelling or machine/deep learning technique. The model allows us to understand text data as tweets or other surveys. Text analysis delivers qualitative and quantitative results with graphs, reports, and tables. Furthermore, from the text analysis, it is possible to extract and obtain specific information, such as keywords, and to categorize survey responses by sentiment (positive, neutral or negative) or specific topic. Based on this aim, the results allow us to obtain trends and patterns.

Text analysis has shown certain advantages:

- Scalable: Several tools allow us to obtain a vast quantity of information from different information surveys (email, news, chats, tweets, comments, social media, etc.)
- Real-time detection: different fields use information by customers or stakeholders or eyewitnesses of specific events. Text analysis is a game-changer in regard to detecting urgent matters 24/7 and in real-time. By training text analysis models to detect expressions and sentiments that imply negativity or urgency, relevant departments can automatically flag tweets, videos, etc., and take action sooner rather than later.
- Consistent Criteria: humans make errors. Training the text analysis model, the algorithms are able to analyse, understand, and sort data much more accurately than humans.

Today, the use of natural language processing is increasing due to substantial improvements in access to data and the increase in computational power. Such an increment allowed practitioners to achieve meaningful results in different fields, such as health care, media, finance, natural disasters mitigation, and human resources. Once natural language processing tools can understand the meaning of a piece of text, and even measure relevant things, several departments can start to prioritize and organize their data in a way that suits their needs.

3.4.1.1 Preprocessing

Some fundamental NLP preprocessing tasks need to be performed before NLP tools can decipher human language. The main drawbacks are the ambiguity and disorganization of the human language. The process of understanding and manipulating involves different steps. A significant number of techniques are applied to reduce the noise of text and to obtain quantitative representation. In natural language processing, the human language is separated into fragments so that the grammatical structure of sentences and the meaning of words can be analysed and understood in context. The result of preprocessing is the transformation of the text into a numeric and organized character. When the text is featured, by number it is possible to apply any traditional statistical or forecasting model.

Convert to lowercase

Within text, the same words can exist with different layouts. To avoid word duplication, such as the different interpretations of the words "*Cat*" and "*cat*", all words change into lower case or upper case characters.

Stemming and Lemmatization

Both techniques standardize words, reducing them to their root forms, but they can remove information.

Stemming refers to cutting the ends of words to correctly achieve this goal most of the time and often involves the removal of derivational units. The final element is known as the *stem*. Stemming is often used for information retrieval to expand search criteria and to reduce the word number for use inside machine/deep learning algorithms.

For instance: *"people" = "people"; "says" = "say"; "troubling" = "trouble"*.

Lemmatization considers the context using a vocabulary and morphological analysis of words to bring the word back to base form. The final element is known as the *lemma*. Lemmatization is often employed for information retrieval to expand search criteria and to reduce the dimensionality of problems in text classification, sentiment analysis and/or topic modelling.

For instance: "payed" = "pay"; "building has floors" = "build have floor"

*I **take** one cat* ⟶ *I **t** one cat* *Stem*

*I **take** one cat* ⟶ *I **took** one cat* *Lemma*

Remove stopwords and words from documents

"Stopwords" are common words or irrelevant characters used in a language, such as articles, prepositions, adverbs, etc. ("the", "a", "to", etc.), and punctuation or special characters ([!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~]). These words or figures do not add information or meaning and are usually removed from texts. After performing all required processes in text processing, some noise is present in the text: too many words with lengths less than 2 or 3 characters. These words should be removed.

Tokenization

Tokenization is a common and fundamental step within natural language processing. Tokenization is a tool for both traditional NLP methods and advanced deep learning within of architectures such as

Transformers. Tokenization is the foremost step when modelling text data, is the process of segmenting text into sentences and is the task of cutting a text into pieces referred tokens. Tokens are the building blocks of natural language and the most common way of processing raw text. Tokens can be either words, characters, or subwords. The following elements can be unique words or words most frequently used and are then used to prepare a vocabulary. Creating a vocabulary is the ultimate goal of tokenization.

For example, this text string*: "There is a landslide, along the way" = There-is-a-landslide-along-the-way.*

There are 3 types of tokenization: word, character and subword (n-gram characters):

- Word tokenization: it is the most commonly employed algorithm. Word tokenization splits a piece of text into individual words based on a certain delimiter (whitespace, comma, etc.). Based on the delimiter, different word-level tokens are formed. There are a few inconveniences, such as out-of-vocabulary words (OVVs), which refer to the new words that are encountered during testing.
- Character tokenization: the text is split into a set of characters. Character tokenization overcomes the inconveniences of word tokenization. The OVV problem is resolved, but the length of the input and output sentences rapidly increases as we are representing a sentence as a sequence of characters, which makes it very difficult to learn the relationship between two characters to form meaningful words. This aspect creates another type of tokenization that falls in between word tokenization and character tokenization.
- Subword tokenization: text is split into subwords (or n-gram characters); for example, the word *lower* can be segmented as *low-er,* and the word *smartest* can be segmented as *smart-est,* etc.

Text encoding

Text encoding is a process in which text is changed into a number/vector representation to preserve the context and relationship between words and sentences (Figure 12). This process allows the machine to understand the pattern associated with any text and can determine the context of sentences.

There are many methods to convert text into numerical vectors:

- Index-Based Encoding: assigns a unique index to all words. All sentences must have the same length. If the sentences do not have the same length, one or more 0s will be added to the end of the shortest sentence.

- Bag of Words: it is the simplest form of text representation in numbers. A sentence can be represented as a bag of words vector or as a string of numbers. First, it is useful for building a vocabulary from all the unique words in the sentence and each word will be marked on the basis of their occurrence.

- TF-IDF encoding: Term frequency–inverse document frequency. Term frequency is the occurrence of the current word in the current sentence with respect to the total number of words in the current sentence.

$$tf_{t,d} = \frac{n_{t,d}}{number\ of\ terms\ in\ the\ document}$$

$n_{t,d}$ = number of times "t" appears in document "d"

Thus, each document and term would have its own term frequency value. Inverse data frequency is the log of the total number of words in the whole data corpus with respect to the total number of sentences containing the current word.

$$idf_t = log\frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}$$

With IDF, the frequency of each word for that particular sentence is included as depending on the number of times a word occurs in a sentence, the TF value can change, whereas the IDF value remains constant, until and unless new sentences are added.

- Word2Vector Encoding: it is a shallow, two-layer artificial neural network that elaborates the text by converting them to numeric "vectorized" words. The input is a word corpus, and the output is a vector space. Therefore, each unique word in the corpus is represented with the generated vector space. Word2Vector encoding is used to reconstruct linguistic contexts of words into numbers. This model captures both syntactic similarities and semantic similarities between two words.

- Transformer Architecture: "Transformer" means the standard encoder and decoder architecture. Their difference is that the decoder part uses masked self-attention with a triangular matrix to prevent tokens from attending their future (right) positions (Qiu et al., 2020).

Figure 12: Evolution of text encoding from the inception of bag-of-words until the modern transformer models, such as BERT, XLM, Roberta etc, that are used today.

3.4.1.2 Modelling-Deep Learning

Deep learning refers to multilayer neural networks in contrast to shallow machine learning (decision trees and support vector machine-SVM). In processing data (Figure 13), deep learning imitates the human brain's neural pathways, building artificial neural networks. The lowest common multiple is the neuron. Traditional machine learning programs linearly work with data analysis, and deep learning's hierarchical function enables a machine to process data using a nonlinear approach.

Currently, advances in learning algorithms and computational performance make deep learning feasible for many complex prediction tasks. In specific fields such as natural language processing, deep learning has shown superior performance with respect to other alternatives of machine learning.

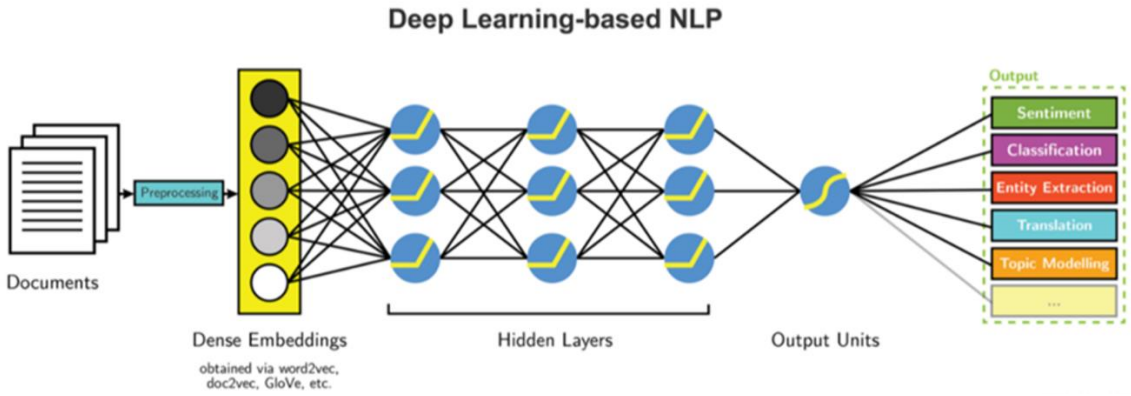

Figure 13: Deep learning approach for NLP (Image credit: https://s3.amazonaws.com/aylien-main/misc/blog/images/nlp-language-dependence-small.png). Deep learning is based on a completely different approach. After an initial preprocessing of raw data, the input is embedded in dense vectors, which can be generated by different techniques, such as word2vec, GloVe and doc2vec. This becomes the new input of the neural network that feeds the hidden layers. Through these layers, the network learns how to reach the goal of the task. It is possible to not specify the language of documents.

With the development of deep learning, various neural networks, such as convolutional neural networks (CNNs) (Kalchbrenner et al., 2014; Kim et al., 2014; Gehring et al., 2017), recurrent neural

networks (RNNs) (Sutskever et al., 2014; Liu et al., 2016), graph-based neural networks (GNNs) (Socher et al., 2013; Tai et al., 2015; Marcheggiani et al., 2018) and attention mechanisms (Bahdanau et al., 2015; Vaswani et al., 2017) have been widely utilized to solve natural language processing tasks, and of consequence, the number of model parameters has rapidly increased. With the development of computational power and deep models, such as Transformers by Vaswani et al., 2017, and the constant enhancement of training skills, the architecture of pretraining has been advanced from shallow to deep. Recently, it has been shown that pretrained models on a large corpus can learn universal language representations. Such evolution is beneficial for NLP tasks and can avoid training a new model from scratch.

Transformers

In 2017, researchers at Google proposed a novel neural network architecture for sequence modelling. Dubbed the *transformer*, this architecture outperformed recurrent neural networks (RNNs) on machine translation tasks, in terms of both quality and training cost. In parallel, an effective transfer learning method named ULMFiT showed that training long short-term memory (LSTM) networks on a very large and diverse corpus could produce state-of-the-art text classifiers with minimal labelled data (Howard et al., 2018). These advances were the catalysts for two of today's most well-known transformers: the generative pretrained transformer (GPT) and bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018). By combining the transformer architecture with unsupervised learning, these models removed the need to train task-specific architectures from scratch and broke almost every benchmark in NLP by a significant margin. Since the release of GPT and BERT, many transformer models have emerged (Figure 14).

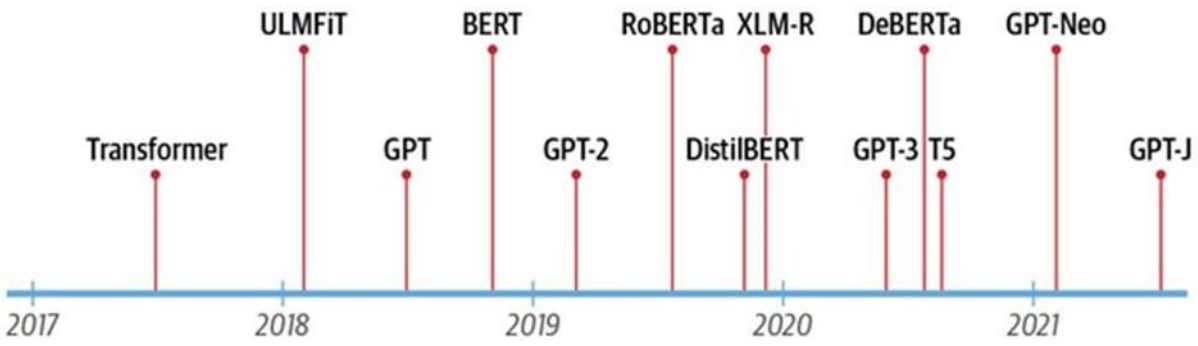

Figure 14: Transformers timeline.

Transformers provide APIs to easily download and train state-of-the-art pretrained models. Using pretrained models can reduce computational costs and save the time involved in training a model from scratch.

Pretraining for transformers model

Pretraining is the act of training a model from scratch: the weights are randomly initialized, and the training starts without any prior knowledge. Pretraining has always been an effective strategy for learning the parameters of deep neural networks, which are fine-tuned on downstream tasks (Qiu et al., 2020). The advantages of pretraining are listed as follows: on large text corpus, it can learn universal language representations; it provides a better model initialization, which usually leads to a better generalization performance and accelerates convergence on the target task; and it can be regarded as a kind of regularization to avoid overfitting on small data (Erhan et al., 2010).
In NLP, pretraining on a large corpus has also been proven to be beneficial for downstream NLP tasks, from shallow word embedding to deep neural models (Qiu et al., 2020). There are two different generations:

1. **The first generation** of pretraining models is pretrained word embeddings. These words represented as dense vectors have a long history (Hinton et al., 1990). The "modern" word embedding was introduced by Bengio et al., (2003) in the pioneering work of the neural network language model.
2. **The second generation** of pretraining models is pretrained contextual encoders. NLP tasks are beyond the word level, and it is important to pre-train the neural encoders on the sentence level or higher. The output vectors of neural encoders are also referred to contextual word embeddings since they represent the word semantics depending on their context (Qiu et al., 2020). Modern pretraining models are usually trained with large-scale corpora and more powerful or deeper architectures (e.g., transformer). Currently, very deep pretraining models have shown their powerful ability in learning universal language representations e.g., OpenAI SPT (Generative Pretraining) (Radford et al., 2018) and Bidirectional Encoder Representation from Transformed (BERT) (Devlin et al., 2018). BERT has become the mainstream approach to adapt pretraining.

Model analysis

The premise is to obtain the implicit linguistic rules and commons sense knowledge hiding in text data, such as lexical meanings, syntactic structures, semantic roles and even pragmatics. The main aim is to describe the meaning of a piece of text in vectors. There are two kinds of word embeddings:

- Non contextual embeddings: Mikolov et al., (2013) demonstrated that words can be represented by a vector. Mikolov et al., (2013), in another analogy study demonstrated that word vectors

produced by Skyp-gram model can capture both syntactic word relationships and semantic word relationships;

for example: vec("China") − vec("Beijing") ≈ vec("Japan") − vec("Tokyo").

The authors determine the compositionality property of word vectors;

for example: vec("Germany") + vec("capital") is similar to vec("Berlin").

- Contextual embeddings: it considers the context-dependent nature of words, and consequently, it is possible to distinguish the semantics of words in different contexts. Given a text x1, x2, ⋯ , xT where each token xt ∈ V is a word or subword, the contextual representation of xt depends on the whole text.

$$[h1, h2, \cdots, hT\ ] = fenc(x1, x2, \cdots, xT\ )$$

where fenc(·) is the neural encoder, and hT is referred to as the contextual embedding or dynamical embedding of token xt because of the contextual information included (Qiu et al., 2020). With this category, BERT is the main task.

Architectures

The transformer architecture follows an encoder-decoder structure (Figure 15) but does not rely on recurrence and convolutions to generate an output.

Transformer encoder: The neural contextual encoders can be classified into two categories:

- Sequence models**:** they capture the local context of a word in sequential order. Sequential models learn the contextual representation of the word with locality bias and find it difficult to capture the long-range interactions between two words. Sequence models are usually easy to train and obtain good results for various NLP tasks (Qiu et al., 2020). Convolutional and recurrent models are mainly employed. Inside the recurrent model, there is bidirectional long short-term memory (LSTM), which is used to collect information from both sides of a word. However, its performance is often affected by the long-term dependency problem.
- Non sequence models: they learn the contextual representation with a predefined tree or graph structure between two words (syntactic structure and semantic relation). A fully connected self-attention model is mainly utilized. A successful example of this model is the transformer; it can

directly model the dependency between every two words in a sequence, which is more powerful and suitable for modelling the long-range dependency of language. This model usually requires a large training corpus and is easy to overfit on small or modest datasets (Radford et al., 2018; Guo et al., 2019).

Currently, the transformer has become the mainstream architecture of pretraining models due to its powerful capacity (Qiu et al., 2020) with transformer encoders and decoders. The encoder receives an input and builds a representation of it (its features), which means that the model is optimized to acquire understanding from the input. These models are often characterized as having "bidirectional" attention and are often referred to as autoencoding models. Once the sentence is transformed into a list of word embeddings, it is fed to the transformer's encoder module. The transformer does not receive one input at a time; it can receive an entire sentence's worth of embedding values and process them in parallel. This approach makes transformers more compute-efficient than their predecessors and enables them to examine the context of the text in both forward and backward sequences. To preserve the sequential nature of the words in a sentence, the transformer applies "positional encoding," which means that it modifies the values of each embedding vector to represent its location in the text. Next, the input is passed to the first encoder block, which processes it through an "attention layer" (Figure 15 on the left). The attention layer tries to capture the relations between two words in the sentence. The attention layer receives a list of word embeddings that represent the values of individual words and produces a list of vectors that represent both individual words and their relations to each other. The output of the attention layer is fed to a feed-forward neural network that transforms it into a vector representation and sends it to the next attention layer. Transformers contain several blocks of attention and feed-forward layers to gradually capture more complicated relationships.
Encoder models are best suited for tasks requiring an understanding of the full sentence, such as sentence classification, named entity recognition (and more general word classification), and extractive question answering.

Transformer decoder: The task of the decoder module is to translate the encoder's attention vector into the output data. The decoder uses the encoder's representation to generate a target sequence, which means that the model is optimized for generating outputs (Figure 15 on the right). At each stage, for a given word, the attention layers can only access the words positioned before it in the sentence. These models are often referred to as autoregressive models and are best suited for tasks involving text generation.

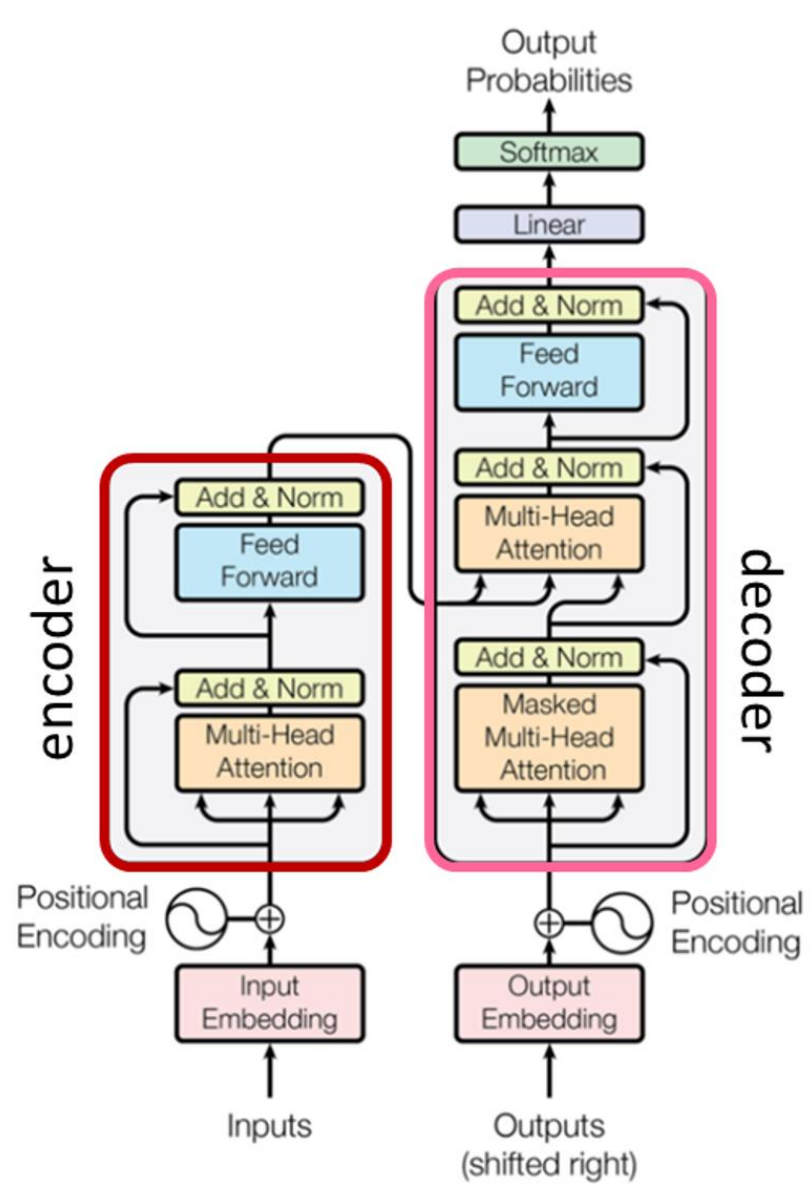

Figure 15: Transformer architecture with an encoder on the left and a decoder on the right.

Each of these parts can be independently used, depending on the task:

- Encoder-only models: Good for tasks that require understanding of the input, such as sentence classification and named entity recognition.
- Decoder-only models: Good for generative tasks such as text generation.
- Encoder-decoder models or sequence-to-sequence models: Good for generative tasks that require an input, such as translation or summarization.

The attention mask can also be utilized in the encoder/decoder to prevent the model from paying attention to certain special words, for instance, the special padding word used to make all the inputs the same length when batching sentences.

3.4.2 Outputs

The outputs of a transformer model (Figure 16), for an analysis of NLP, make it possible to represent and extract information from text in manner a qualitative and quantitative manner. On the basis of data sources, it is possible to obtain specific information:

- Text: text classification, information extraction, question answering, summarization, translation, and text generation in more than 100 languages.
- Images: image classification, object detection, and segmentation.
- Audio: speech recognition and audio classification.
- Multimodal: table question answering, optical character recognition, information extraction from scanned documents, video classification, and visual question answering.

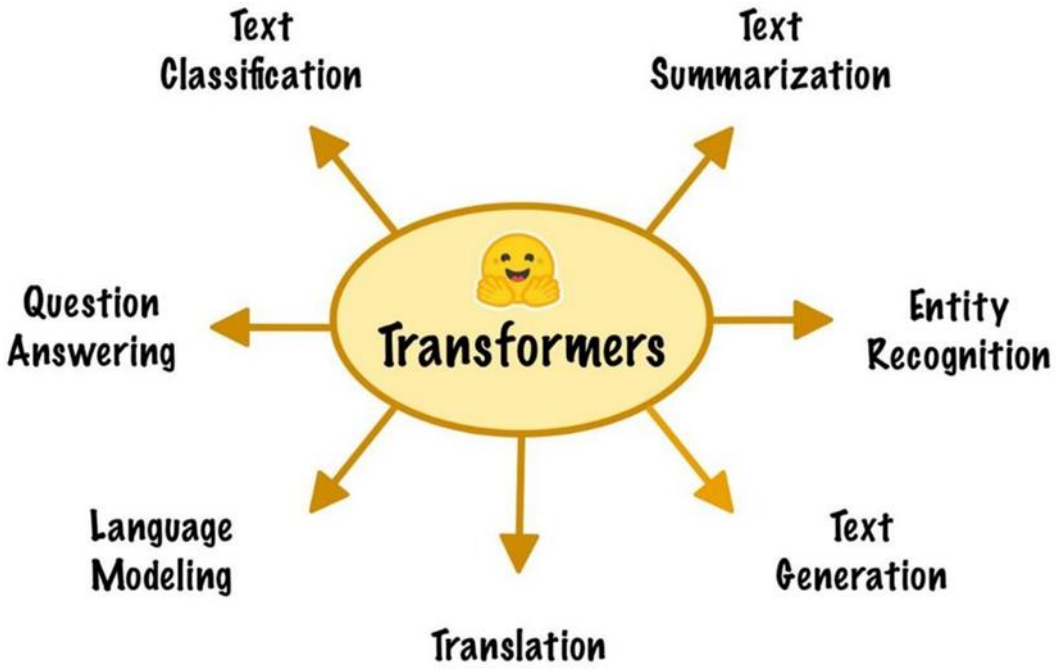

Figure 16: Different outputs can be obtained using transformer architecture. Possible outputs can be: text summarization, entity recognition or name entity recognition (NER), text generation, translation from one language to another language, language modelling, question answering and text classification.

A wide range of NLP use cases exist, and some examples of algorithms are presented as follows:

Classification: This is the process of assigning predefined tags or categories to unstructured text. Classification is versatile and can organize, structure and categorize any form of text to deliver meaningful data and solve problems. This process is considered one of the most useful natural

language processing techniques. The most common text classification tasks are sentiment analysis, topic modelling, language detection and intent detection.

- Sentiment analysis: consists of the automated processing of texts to identify and classify subjective information related to sentiments. This information might be an opinion, a judgement, or a feeling about a particular topic or product feature. The common type of sentiment analysis is '*polarity detection'* and involves classifying statements as positive, negative or neutral.
- Topic modelling: common example of text classification that organizes text by subject or theme. One example is the latent Dirichlet allocation (LDA) model. This relatively new algorithm (invented less than 20 years ago) works as an unsupervised learning method that discovers different topics underlying a collection of documents. In unsupervised learning methods such as this method, there is no output variable to guide the learning process, and data are explored by algorithms to identify patterns.

Text extraction: Text analysis techniques extract pieces of data that already exist within any given text. It is possible to obtain several pieces of information, such as keywords and entities.

- Keywords: they are the most used and most relevant terms within a text, words and phrases that summarize the contents of the text.
- Word frequency: this is a text analysis technique that measures the most frequently occurring words in a given text using numerical statistics.
- Co-occurrence: Given a corpus of documents, a co-occurrence network is an undirected graph, with nodes corresponding to unique words in a vocabulary and edges corresponding to the frequency of words co-occurring in a document. You can use a co-occurrence network to discover which words commonly appear with a specified word. A word cloud is a similar approach. Both are methods of qualitative representation.
- Name Entities Recognition (NER)**:** the entities are the most important objects of a particular sentence, as noun phrases, verbs or both. NER can automatically scan entire articles and to obtain more fundamental entities in a text and classify them into specific categories. The categories can be people's name, company name, geographic locations, dates and times, names of events and organizations.

3.4.3 Bidirectional Encoder Representations from Transformers (BERT)

The Transformer Architecture was released in December 2017 in a Google machine translation paper titled “Attention Is All You Need” (Vaswani et al., 2017). That paper tried to obtain models that could automatically translate multilingual text. The attention mechanism, which highlights the important information from the contextual information by setting different weights, has also been applied to improve accuracy (Zhang et al., 2018). The more important event is the introduction of bidirectional transformers for language understanding (Devlin et al., 2018). Bidirectional Encoder Representations from Transformers (BERT) was first released in October 2018 in “*Pre-Training of Deep Bidirectional Transformer for Language Understanding*”. BERT was pretrained on a massive unlabelled text corpus comprising whole Wikipedia (~2.5B words) text data and Google’s BooksCorpus (~800 M words). BERT’s training was made possible due to the novel transformer architecture and accelerated by using tensor processing units (TPUs - Google’s custom circuit built specifically for large ML models). Sixty-four TPUs trained BERT over the course of 4 days. BERT utilizes a transformer to create the vector representation (Dharma et al., 2022). The attention mechanism of the transformer architecture allows models such as BERT to bidirectionally process text by:

1. Allowing parallel processing: Transformer based models can process text in parallel, avoiding the sequential processing of text.
2. Storing the position of the input: the transformer architecture directly encodes the position of the word into the embedding. This is a “marker” that lets attention layers in the model identify the location of the word or text sequence that they are viewing. This trick means that these models can keep processing sequences of text in parallel, in large volumes with different lengths, and still know exactly in what order they occur in the sentence.
3. Making lookup easy: Transformer based models can simply look up any word in a sentence at any time.

BERT does not use the decoder of the transformer architecture (Figure 17).

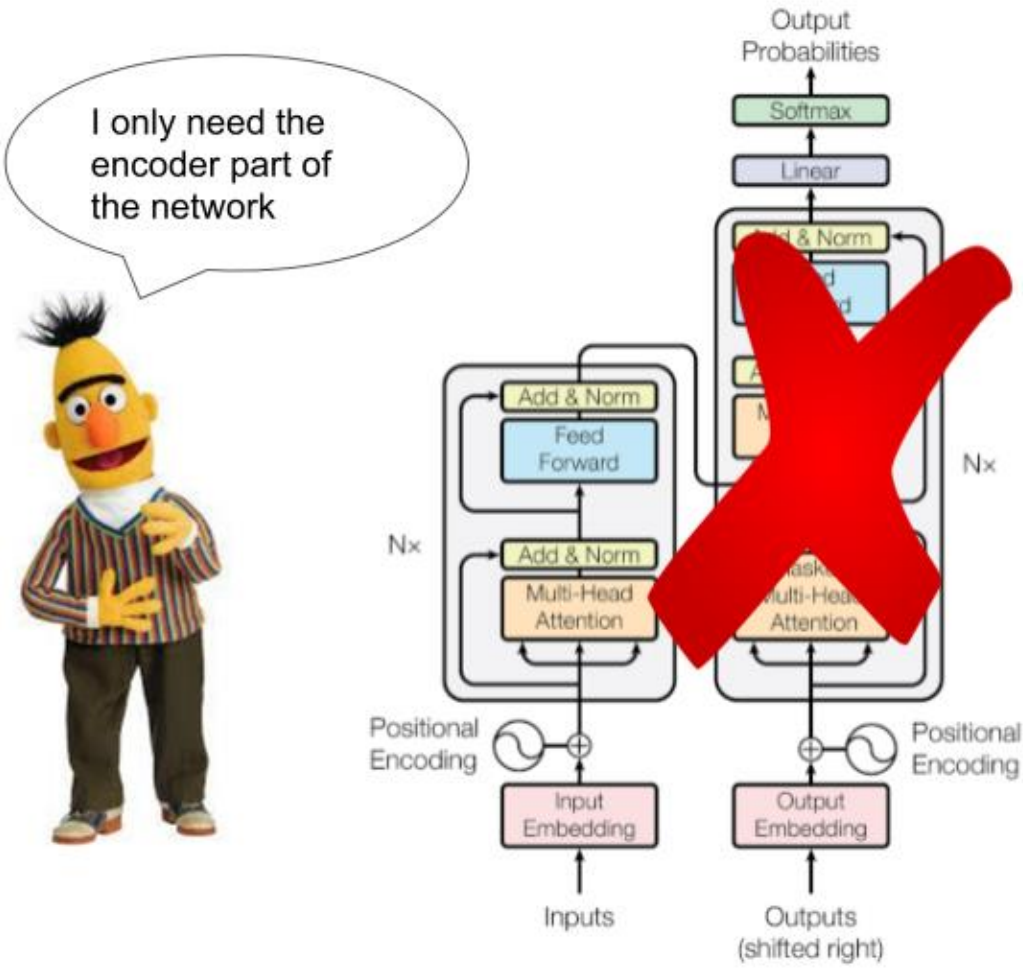

Figure 17: Transformer architecture within BERT. Only the encoder part is present.

BERT uses the encoder part of the transformer. As a result, using the encoder enables BERT to encode the semantic and syntactic information in the embedding, which is needed for a wide range of tasks. Using only the encoder BERT is not designed for tasks such as text generation or translations. BERT can be trained on multiple languages, but it is not a machine translation model. Therefore, the output of BERT is an embedding, not a textual output. In contrast, if the decoder is employed, the output would be a text, which could be directly applied without needing to perform any further actions. BERT takes the output of the encoder and uses it with training layers that perform two innovative training techniques. First, BERT proposes a masked language model (MLM) inspired by the Cloze task (Taylor, 1953), in which 15% of the input tokens are randomly masked by a special label [mask] and then those masked tokens are predicted. Second, BERT introduces next sentence prediction (NSP) to the training process (Zhou et al., 2022). These are ways to unlock the information contained in the BERT embeddings to obtain the models to learn more information from the input. In this way, BERT forces the encoder to try and “learn” more information about the surrounding text to better predict the

hidden or “masked” word. Then, for the second training technique, it obtains the encoder to predict an entire sentence given the preceding sentence.

BERT can be employed on a wide variety of common language tasks: sentiment analysis, question answering, text prediction, text generation, summarization etc.

BERT is available in two sizes: BERT-BASE (containing 12 transformer layers and 768 hidden layers) and BERT-LARGE (containing 24 transformer layers and 1024 hidden layers) (Lee et al., 2022). Table 8 shows the architecture of each.

	Transformer Layers	Hidden Size	Attention heads	Parameters	Processing	Length of training
BERT-BASE	12	768	12	110 M	4 TPUs	4 days
BERT-LARGE	24	1024	16	340 M	16 TPUs	4 days

Table 8: **Transformer Layers**: Number of transformer blocks. A transformer block transforms a sequence of word representations into a sequence of contextualized words (numbered representations); **Hidden Size**: Layers of mathematical functions, located between input and output, that assign weights (to words) to produce a desired result; **Attention Heads**: the size of a transformer block; **Parameters**: Number of learnable variables/values available for the model; **Processing**: Type of processing unit used to train the model; **Length of Training**: Time it took to train the model.

Multilingual Pretraining Using BERT

However, most of these BERT-based models are based on an English-centric design; consequently, researchers have attempted to develop models based on languages other than English (Lee et al., 2022).

Multilingual BERT (Yang et al., 2019) retains the model structures of BERT but replaces the pretrained corpus attributes with those that include more than 100 languages. This approach results in significant performance improvements over the original BERT in natural language comprehension tasks. However, this finding that its vocabulary size is large and that its size inefficiently increases owing to the processing of more than 100 languages, which consequently restricts memory efficiency (Lee et al., 2022).

Cross-lingual modelling (Lample et al., 2019) trains a model via unsupervised-learning-based pretraining, where continuous learning in English and other languages is simultaneously applied. BERT was pretrained using a dataset containing 100 languages. This method significantly improves symbolic performance in multilingual tasks other than those in English. However, cross-lingual modelling is not comparable to well-preprocessed models in English in terms of accuracy (Lee et al., 2022).

RoBERTa is a novel and improved recipe for training BERT models that can match or surpass many post-BERT methods (Liu et al., 2019). RoBERTa optimizes the training process and offers the training process

more time, with larger batch sizes and more data (Zhou et al., 2022). RoBERTa is an unsupervised model that relies only on monolingual data.

A further step forward is the development of transformer multilingual. XLM-RoBERTa is a multilingual version of RoBERTa, where XLM is an acronym for the cross-lingual language model. XLM-RoBERTa is pretrained as a masked language model on 100 languages and 2.5 TB of filtered common crawl data (Conneau et al., 2020).

3.4.3.1 BERT for Text Classification

Text classification is referred to as extracting features from raw text data and predicting the categories of text data based on such features. The general architecture of deep learning is shown in Figure 18. BERT takes an input of a sequence with a maximum of 512 tokens and outputs the representation of the sequence (Figure 18 -Tokenization). The sequence has one or two segments in which the first token of the sequence is always [CLS] which contains the special classification embedding and another special token [SEP] is used for separating segments (Sun et al., 2019) (Figure 18 - Deep Learning). Each token is converted into fixed-size word vector, also known as word embedding. Within Deep learning is applied the Droupt layer. Droupt is a technique used to ignore the randomly selected neurons from deep learning models and it reduces over-fitting (Madichetty et al., 2021).

For text classification tasks, BERT takes the final hidden state T_n of the first token [CLS] as the representation of the whole sequence. A simple softmax classifier is added to the top of BERT to predict the probability of class label (Figure 18 - Prediction).

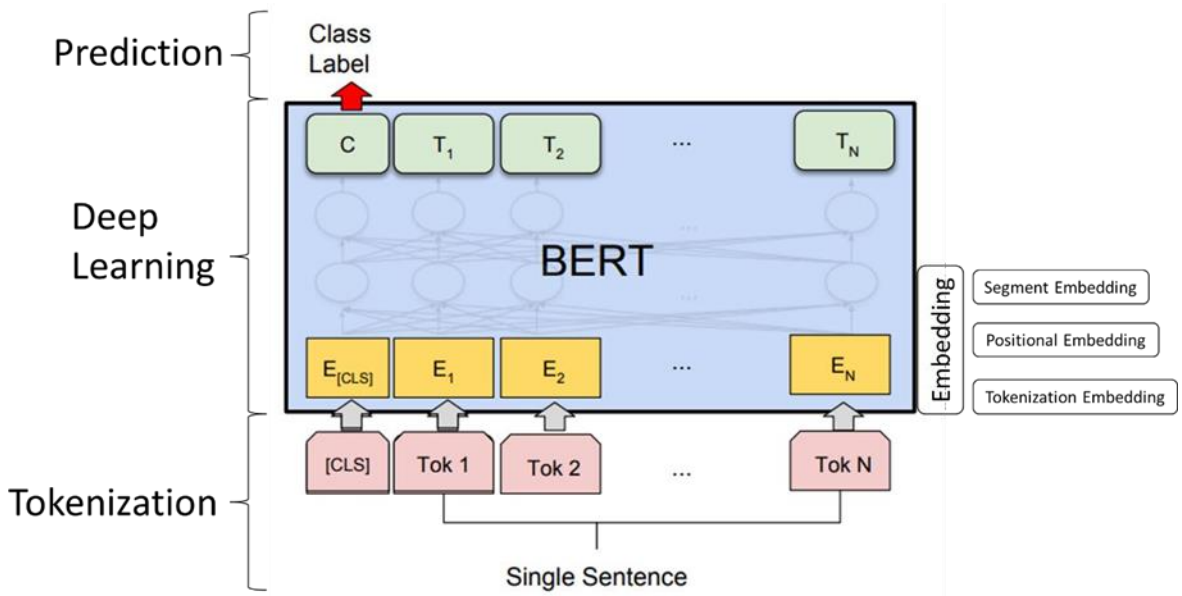

Figure 18: Illustrations of Fine-tuning BERT by Devlin et al., 2018, modified with specifics about tokenization, deep learning with an embedding part and the class label prediction. Word embedding is the conversion of words in a document to vectors in which values assigned to them are closer in the vector space.

Literature review

Most studios used the English language for applying deep learning techniques. Numerous deep learning models with BERT architecture have been proposed in these few years for text classification of tweet text, as shown in Table 9.

Studios were placed in chronological order and the topics of application range from sentiment analysis techniques (Alaparthi et al., 2020; Geetha et al., 2021) to event detection (Jain et al., 2019; Madichetty et al., 2020; Yumeng Hu et al., 2021; Liu et al., 2021; Huang et al., 2022; Dharma et al., 2022; Zhou et al., 2022) using different languages.

Alaparthi et al., (2020) and Geetha et al., (2021) use the BERT classifier to determine the sentiment positive or negative of users. Alaparthi et al., (2020) used film reviews as dataset for the classification of sentiment. They demonstrated the remarkable classification accuracy of the BERT method, reaching 92%. Geetha et al., (2021) performed the same sentiment analysis using Amazon product dataset. They compared the BERT model to other machine learning models such as Naevi Bayesian and Support Vector Machine (SVM). The results demonstrated the remarkable classification capability of a transformer architecture. The accuracy was 88,48%. Textual analysis of vocabulary, words, grammar, and other features such as text sentiment provide information that can be leveraged in the post-event environment (Li et al., 2021). Such textual analysis is now widely used in sociology, psychology, marketing, and elsewhere to draw conclusions from what appears to be relatively descriptive and qualitative information (Das et al., 2019; Mahoney et al., 2019; Majumdar & Bose, 2019; Osorio-Arjona & García-Palomares, 2019; Plunz et al., 2019; Reboredo & Ugolini, 2018).

Much of the recent research on social media and disasters capturing textual analysis converges on themes of early warning, emergency response, and behaviour analysis (Li et al., 2021). Spruce et al., (2020) leveraged sentiment analysis to measure the impact of storms and other extreme weather events. Disaster tweet classification study can be considered as a Natural Language Processing (NLP) task. The use of a deep learning model has started to become more common for natural language processing tasks. Various methodologies were applied for classifying crises, either using the BERT classifier alone or incorporating the classifier with other deep-learning techniques. Jain et al., (2019), Liu et al., (2021), and Zhou et al., (2022) used the BERT method alone to check whether the text describes a crisis or not. The first two authors demonstrated the effectiveness of the classifier for flood, hurricane and earthquake events. Both studies used a binary classification (0 and 1) estimating a maximum accuracy of 95% with Liu et al., (2021). Zhou et al., (2022), use the BERT method for recovery activities after a natural event, classifying text according to aid information, complete interridge and victims. They chose original English tweets containing the 5-digit zip code of coastal Texas as potential rescue request tweets. This study for each label compares a different model, also implementing other deep learning techniques such as convolutional neural networks (CNN), or Long short-term memory (LSTM) in the BERT model.

Madichetty et al., (2020), Yumeng Hu et al., (2021), Huang et al., (2022) and Dharma et al., (2022) implemented other deep learning models to the BERT model to improve text classification. In particular, the implementation of CNN with BERT embedding is predominant. Madichetty et al (2020) used this combination to delineate whether the text is informative with respect to a crisis or not. The datasets used for analysis were part of different events that occurred in certain parts of the world: the typhoon in Hagupit (Philippines), explosion in Hyderabad (southern India) and the shooting at Sandy Hook (Connecticut-USA). The dataset features tweets in Hind and English language. Different embedding models were used, reaching a maximum accuracy of 96%. Similar analysis was conducted by Dharma et al., (2022). This study uses extracted Twitter data for different natural events: earthquake, flood, pyroclastic flows, eruption, tsunami, drought, landslide, typhoon and others. These data constitute a single dataset in Indonesian language. The data were manually classified with a Boolean value in which the value 1 is distributed to the data with disaster and a non-disaster was classified with 0. CNN with pre-trained BERT embedding was able to get the best result (Dharma et al., 2022). The accuracy was 97,16%. At the same time, Huang et al., (2022) proposed an integrated approach to detect all four kinds of emergency events early, including natural disasters, man-made accidents, public health events (COVID-19), and social security events. For text classification, massive Weibo posts in Chinese language were used to train different models. The classification phase uses the integrated approach combining BERT and an attention-based bidirectional long short-term memory model (BERT-Att-BiLSTM) to detect emergency-related posts. The highest accuracy was of 90,58%.

Yumeng Hu et al., (2021) conducted a more complex study on the use of deep learning techniques for seismic P-wave detection (positive or negative). TransQuake is an advanced deep learning approach from Transformer. TransQuake exploits the STA/LTA algorithm to fit the three-component structure of seismic waves as input and exploits the multi-headed attention mechanism to conduct pattern learning (Yumeng Hu et al., 2021). At the same time, Sánchez et al., (2022) introduced a new multilingual and multi-domain crisis dataset, containing 53 crisis events and more than 160.000 messages. They proposed an empirical transfer learning, using crisis data from high resource language (as English) to classify data from other languages (as Italian, Spanish and French). The authors used different model for binary classification of tweets that are related and unrelated to crisis. Considering Italian language, the authors considered data from flood events (on Sardinia and Genova events) and earthquake events (on L'Aquila events). Both datasets come from SoSItalyT4 by Cresci et al., (2015). The best performance, for flood event, was archived in the Cross-lingual & Multi-domain scenario, using XLM-RoBERTa model. In this scenario, the training was featured by multiple domains in one language (e.g. floods, earthquakes and hurricane in English) while the test set was characterized by a new event in another language (e.g. flood in Italian). The highest F1 value was of 0,84.

A high performance was achieved using Multilingual & Multi-domain scenario with Machine Translation (MT)+BERT for earthquake event. In this case the model was trained with English and Italian tweets about floods, earthquakes and hurricanes to then classify earthquake-related messages in Italian. The best value of F1 was of 0.82.

Authors	**Target**	**Event**	**Language**	**Model**	**Results**	
Jain et al., (2019)	Relevant to crisis or no	Earthquake	English	BERT	P	0,83
		Flood			R	0,76
		Hurricane			F1	0,78
					A	0,76
Alaparthi et al., (2020)	Sentiment	IMDB film reviews	English	BERT	P	0,92
					R	0,92
					F1	0,92
					A	0,92
Madichetty et al., (2020)	Relevant to crisis or no	Hurricane	English	CNN+BERT embedding-Large	P	0,95
					R	0,98
		Explosion		CNN+BERT embedding-Base	F1	0,97
					A	96%
Liu et al., (2021)	Relevant to crisis or no	Crisislex	English	DistilBERT	F1	0,95
					A	95%
Hu et al., (2021)	Positive or Negative	Waves P		Transoformer+CNN+ STA/LTA	P	0,71
					R	0,67
					F1	0,68
					A	0,95

Geetha et al., (2021)	Sentiment	Review from Amazon	English	BERT-Base-Uncased model	P	0,88
					R	0,86
					F1	0,89
					A	0,88
Huang et al., (2022)	Crisis classification	Natural disaster, accident, public health event, social security event	English	BERT	P	0,84
					R	0,89
					F1	0,86
					A	0,90
				BERT-Att-BiLSTM	P	0,85
					R	0,93
					F1	0,89
					A	0,90
Dharma et al., (2022)	Disaster yes or no	Earthquake, Flood, Pyroclastic flow, Drought, Typhoon, Tsunami, Landslide, Eruption	Indonesian	CNN+BERT embedding	P	0,97
					R	0,96
					F1	0,97
					A	0,97
Zhou et al., (2022)	Help or no help	Hurricane	English	BERT-CNN	P	0,89
					R	0,93
					F1	0,9
Sánchez et al., (2022)	Relevant to crisis or no	Flood	Italian	XLM-RoBERTa	F1	0,84
		Earthquake		MT-BERT	F1	0,82

Table 9: State-of-the-art applications of BERT for text classification tasks. Different aims have been presented from sentiment analysis to crisis classification, or consequently, with help messages or victim information. The best results of the classification test, combining two or more methods, such as CNN+BERT, have been shown. Results are rappresented as metrics, P= precision, R= recall, F1= F1 and A= accuracy.

3.4.4 Methodological BERT for landslide events

In this work, the tweet database underwent manual classification based on relevance and localization. This classification allows us to identify the most relevant tweets in terms of the temporal and spatial accuracy of landslide event identification. Furthermore, different coordinates have been attributed to tweets on the basis of text and then checked in the real location. The classification features by 2 classes for each label: landslide 0 and 1 (“tweet text no describes or describes one landslide event”); coordinate 0 and 1 (“tweet text no identifying or identifying one location”) (Table 10).

Tweet text	Landslide yes or not	Coordinate
#SanremoNews Frana sulla Statale 20 ad Airole: questa mattina riunione tra i Sindaci in Prefettura http://t.co/B1mwrMLq	1	1
passati un pò prima del disastro nelle zone colpite dall'alluvione..e nel tratto della frana..circa un'ora prima..Ho i brividi.#alluvione	1	0
Vodafone avvisa: "Servizio ripristinato, disagi per maltempo e crollo viadotto" https://t.co/YA6a9ydDPQ	0	0

Table 10: Examples of manual classification for tweet text considering information about landslide and coordinated.

Figure 19 shows several steps to obtain the classification for each tweet text using deep learning with architecture transformers to obtain information about landslide events.

The manually classified database was preprocessed to remove special characters. By applying a supervised deep learning technique, it is possible, in this case, to obtain labelled results 0 and 1.

Recently, transformer-based pretrained language models have demonstrated stellar performance in natural language tasks. Bidirectional encoder representations from transformers (BERT) have achieved outstanding performance (Lee et al.,2022).

In this work, XLM-RoBERTa (a Cross Lingual Model) was chosen as the method with architecture transformers (Table 11) applied to tweet data with Bayesian classification (or binary classification 0 and 1). XLM-RoBERTa is a multilingual model trained on 100 different languages. Unlike some XLM multilingual models, it does not require language tensors to understand which language is used and should be able to determine the correct language from the input ids. XLM-R has been chosen for two reasons: i) multilingual models can outperform their monolingual BERT counterparts (Conneau et al., 2019); ii) there is a proliferation of non-English models, and the multilingual models are a helpful compromise to classify text in non-English language.

The complete database was preprocessed, and the model 'm-polignano-uniba/bert_uncased_L-12_H-768_A-12_italian_alb3rt0' (Polignano et al., 2019) was chosen for tokenization.

Finally, the XLMRobertaForSequenceClassification method was applied with the 'xlm-roberta-base' model to predict 2 number labels (0 and 1). Table 11 shows several aspects of the XLM-RoBERTa method and its linked parameters.

Architecture	Method	Model	Transformer Layers	Hidden Size	Attention heads	Parameters	vocab size
BERT	XLMRobertaForSequenceClassification	xlm-roberta-base	12	768	12	250M	250k

Table 11: **Transformer Layers**: Number of transformer blocks. A transformer block transforms a sequence of word representations into a sequence of contextualized words (numbered representations); **Hidden Size**: Layers of mathematical functions, located between the input and output, that assign weights (to words) to produce a desired result; **Attention Heads**: The size of a transformer block; **Parameters**: Number of learnable variables/values available for the model; **Vocab size**: vocabulary for text analysis in many languages.

Three types of preprocessing were applied to obtain the best results with XLM-RoBERTa. The first preprocessing considered the dataset without cleaning; the second considered all possible parameters for removing, and the third only removed some parameters within the text. Figure 19 illustrates the workflow of the work. From the database without labels, preprocessing was applied before deep learning for classification. The result is to obtain a text classification on the basis of landslide information.

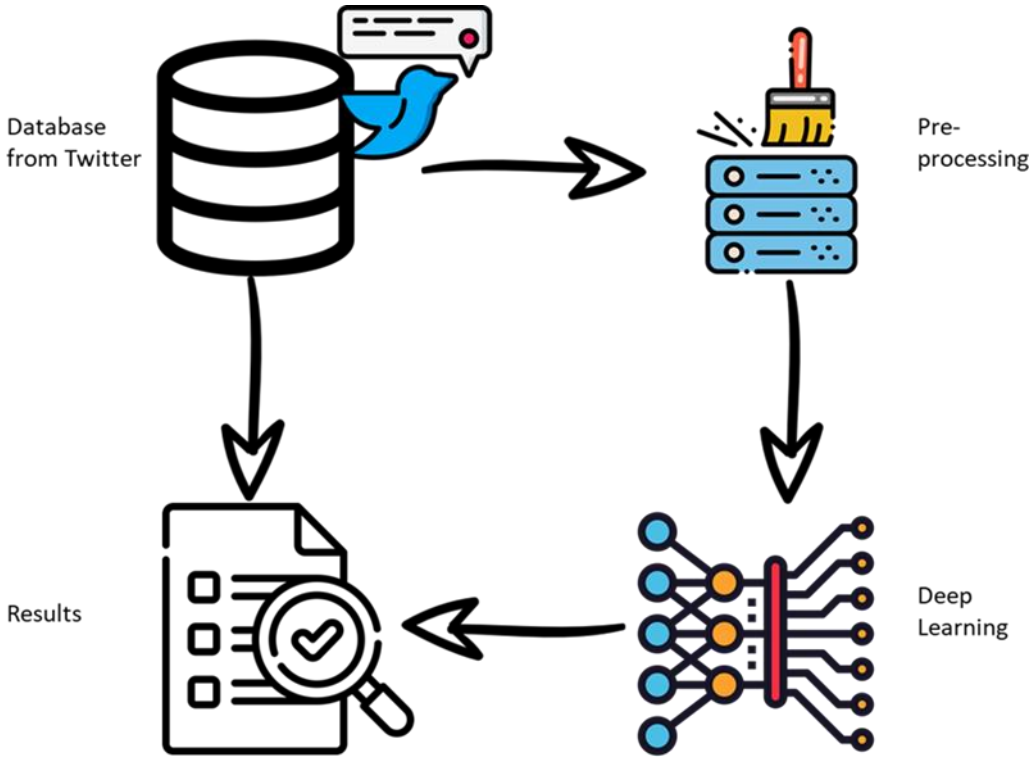

Figure 19: Workflow of tweet analysis with preprocessing, deep learning and result with one tag, 0 or 1 for each tweet text.

Parameter setting

The datasets were randomly divided into 80% training and 20% testing. The training dataset was further randomly divided by 20%, resulting in the validation dataset. This operation was carried out

for each of the three tests, so only the test set was kept constant, changing the training and validation datasets each time.

The three tests maintained the following parameters (Table 12):

- **Maximum length:** text beyond which it is truncated;
- **Batch size**: the number of samples processed before updating the model;
- **Epoch**: Epoch indicates the number of passes of the entire training dataset the deep learning algorithm has completed. An epoch comprises one or more batches;
- **Seed**: the randomness of an artificial neural network (ANN) is when the same neural network is trained on the same data, and it produces different results. This randomness in the results makes the neural network unstable and unreliable. To make the randomness predictable, we use the concept of *seed.* Seed helps obtain predictable, repeatable results every time;
- **Learning rate (Adam): Adam** is an adaptive learning rate optimization algorithm that has been designed specifically for training deep neural networks. The first value is referred to as the learning rate or step size. Larger values (e.g., 0,3) result in faster initial learning before the rate is updated. Smaller values (e.g., 1,0e-5) slow learning right down during training. The second value, **Epsilon,** is a very small number to prevent any division by zero in the implementation (e.g., 10e-8).
- **Early stopping**: to avoid continuous iterations, an early stopping has been set. Hence, if the model does not improve after total tries, it stops.

Max length	Batch size	epoch	seed	Learning rate	epsilon	Early stopping
128	32	100	45	2e-5	1e-8	15

Table 12: Table with parameter settings for the model. Some values have been retrieved from the state-of-the-art and retained constants (such as Seed, Learning rate, and Epsilon).

The metrics used to evaluate the performance of the model were accuracy, confusion matrix, precision, recall, F1 score and the calculation of the AUC with the rock curve.

- **Accuracy** is the ratio of the number of correct predictions to the total number of input samples.

$$accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions\ made}$$

- **The confusion matrix** describes the complete performance of the model. TruePositive (TP) and TrueNegative (TN) are data predicted correctly, in contrast to FalseNegative (FN) and FalsePositive (FP).

Real: no or 0	TN	FP
Real: yes or 1	FN	TP
	Predicted: no or 0	Predicted: yes or 1

- **Precision** is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$precision = \frac{TruePositives}{TrueProsition + FalsePositives}$$

- **Recall** is the number of correct positive results divided by the number of *all* relevant samples.

$$recall = \frac{TruePositives}{TrueProsition + FalseNegatives}$$

- The **F1 score** is the harmonic mean between precision and recall. The range is [0, 1]. This parameter describes the precision of the classification and its robustness.

$$F1\ score = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

- **The receiver operating characteristic (ROC) and area under the curve (AUC)** of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. ROC is a probability curve, and AUC represents the degree of separability. The higher the AUC is, the better the model is at predicting. The ROC curve is plotted with TruePositiveRate (TPR) or sensitivity against FalsePositiveRate (FPR).

$$TruePositiveRate\ (TPR) = \frac{TruePositive}{FalseNegative + TruePositive}$$

$$FalsePositiveRate\ (FPR) = \frac{FalsePositive}{TrueNegative + FalsePositive}$$

Library for Python for deep learning

To apply deep learning, some libraries have been used inside Python. The combination of libraries and the many tools for NLP analysis, such as PyTorch, has made Python one of the most preferred programming languages for performing text analysis.

Below, two main libraries are listed.

PyTorch is a new entry within a deep learning framework based on Torch. It was developed by Facebook's AI research group and open-sourced on GitHub in 2017. It is utilized for natural language processing applications. PyTorch is simple, easy to use, flexible and efficient for memory usage and a dynamic computational graph. It has a complex architecture, and its readability is lower than that of other packages (e.g., Keras).

Transformer architecture was introduced in June 2017. The focus of the original research was on translation tasks. All transformer models were trained as language models. This means they have been trained on large amounts of raw text in a self-supervised fashion. This type of model develops a statistical understanding of the language it has been trained on but it is not very useful for specific practical tasks. Its goal is to provide a single API through which any transformer model can be loaded, trained and saved. The library's main features are as follows:

- Ease of use: downloading, loading and using a state-of-the-art NLP model for inference can be done in just two lines of code.
- Flexibility: at their core, all models are simple PyTorch nn.Module classes and can be handled as with any other model in their respective machine learning (ML) frameworks.
- Simplicity: hardly any abstractions are made across the library. The "all in one file" is a core concept: a model's forward pass is entirely defined in a single file so that the code itself is understandable and hackable.

Transformers provide several models already pretrained, such as XLM-RoBERTaTokenization, XLM-RobertaModel and XLM-RoBERTaForSequenceClassification.

4 RESULTS

The results are shown below. The first analyses were carried out within the state-of-the-art of data mining. The state-of-the-art involved the classification of two datasets from the multi-risk information gateway (MIG) platform developed by Battistini et al., 2013,2018. Two datasets harvested newspaper articles from Google News for two types of events, landslides and floods. The period of analysis started from 2010 until 2019. The classified datasets constitute a valid and additional inventory of landslide and flood phenomena throughout Italy. From the classification, various analyses were carried out to determine the spatial and temporal distribution of the events. Moreover, the different nomenclature attributed to the different subdivisions of the dataset made it possible to make further considerations in terms of hazard, media impact and temporal distribution. Due to the different manipulations, various correlations with other existing data sources (rainfall data, ReNDiS and Polaris) were applied. More targeted analyses have been carried out for landslide events, also considering three types of hazard maps (landslide hazard, building at risk and people at risk) from ISPRA.

In parallel, the classified database was subjected to some natural language processing techniques to obtain various information, including keywords. The keywords allowed us to take that extra step in the data mining technique. The data mining technique has thus far been applied to newspaper news, but now, with the appropriate use of keywords, also within the Twitter dashboard. Various data slots, which form a single database, were extracted from 2011 to 2019. The choice of periods was based on the temporal distribution of landslide events from newspaper reports. As before, this database was manually classified according to the landslide information and the presence of text and the actual coordinates of the event.

In addition, to validate the tweet dataset, several slots have been compared to classified news.

Some case studies of landslide events are presented considering different triggers (riverbank erosion and subsequent heavy rainfall) or the presence of victims. More analysis has been applied to the landslide event in Liguria that occurred in November 2019. Tweets were compared with rainfall data. The first landslide tweets were recorded 19 minutes after the event. The timeliness and spread of the publications demonstrated that Twitter is a valuable source of information for natural events.

The binary classified dataset based on landslide information provided a solid basis for applying deep learning techniques. Three different preprocessing steps were applied. Multiple tests allowed us to

obtain a model that better interprets and classifies the textual data. Moreover, to validate the capability of the model, it has been tested on the new 2020 slot. This dataset has been compared with the classified news dataset. Another validation has been applied considering the case study of Liguria.

4.1 Newspaper article about landslides and floods

The news database was classified into three classes based on news relevance, localization accuracy and time of publication to distinguish "News referred to recent events" (Class 1) from "News generically referred to events" (Class 2) and "News not related to event" (Class 3). This classification allowed us to define, at the national scale, the areas and periods mainly involved in landslide and flood events. Figure 20A shows the three classes within the Landslide database, while Figure 20B shows two classes in the Flood database. Within flood database were not detected articles in class three.

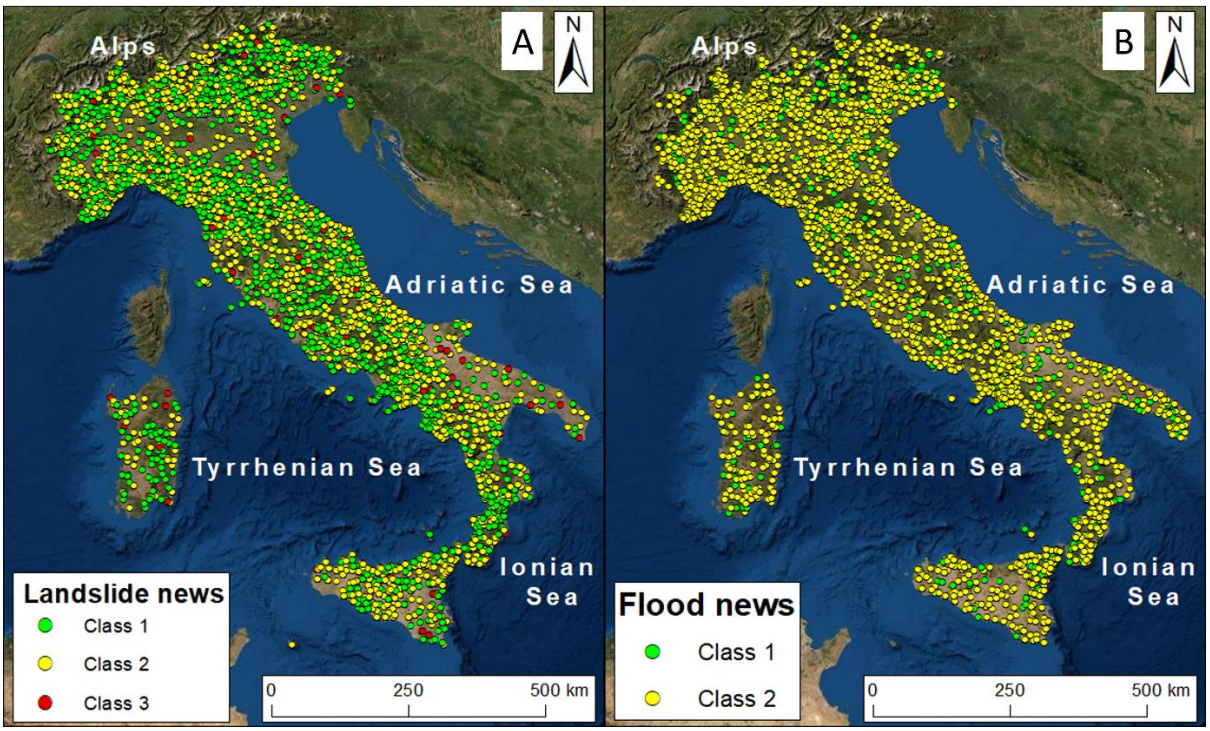

Figure 20: General distribution with the classification of the news used in Italy for landslides in **A** and flood events in **B**. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

4.1.1 Landslide news

Spatial distribution

The data mining algorithm used cannot identify the exact location of a landslide, since it is not usually reported in newspapers; therefore, the data have been grouped on a regional base (Figure 21A) and on a provincial base (Figure 21B) to identify the areas with a higher number of landslide news. Class 2 news has been used only on a regional scale aggregation since some of them do not provide adequate localization accuracy for a more detailed analysis. According to the spatial distribution of the news, during the last 10 years, 41,7% of the municipalities suffered at least one landslide.

The regions most affected by landslides are mainly in the northern part of the country. Liguria and Lombardia are the regions with the highest number of news (classes 1 and 2) and therefore of article publication (articles referring to the same landslide event are grouped into a single “Landslide news”). For example, Liguria has 36.451 articles referring to 4318 landslide news (classes 1 and 2, Figure 21A); among them, 19.844 articles refer to 1174 recent “Landslide events”, and in particular, Genova is the province most affected by landslides (Figure 21B).

In addition to the alpine area, several other provinces over the country showed a relevant number of news (Salerno, Messina, Savona and Sondrio), and they are mainly located along the western coast (Tyrrhenian seacoast) and along the Apennines mountain belt (Figure 21B), which is historically affected by landslides because of its geological origin and the high frequency of clayey slopes.

The Puglia region (Figure 21A) and the provinces along the northeast coast (Figure 21B) show a lower number of landslide news because they are mainly flat areas, and fewer landslides are obviously expected (Figure 21A); this is true as well for the southern part of Lombardia and Veneto and the north-eastern part of the Emilia-Romagna region.

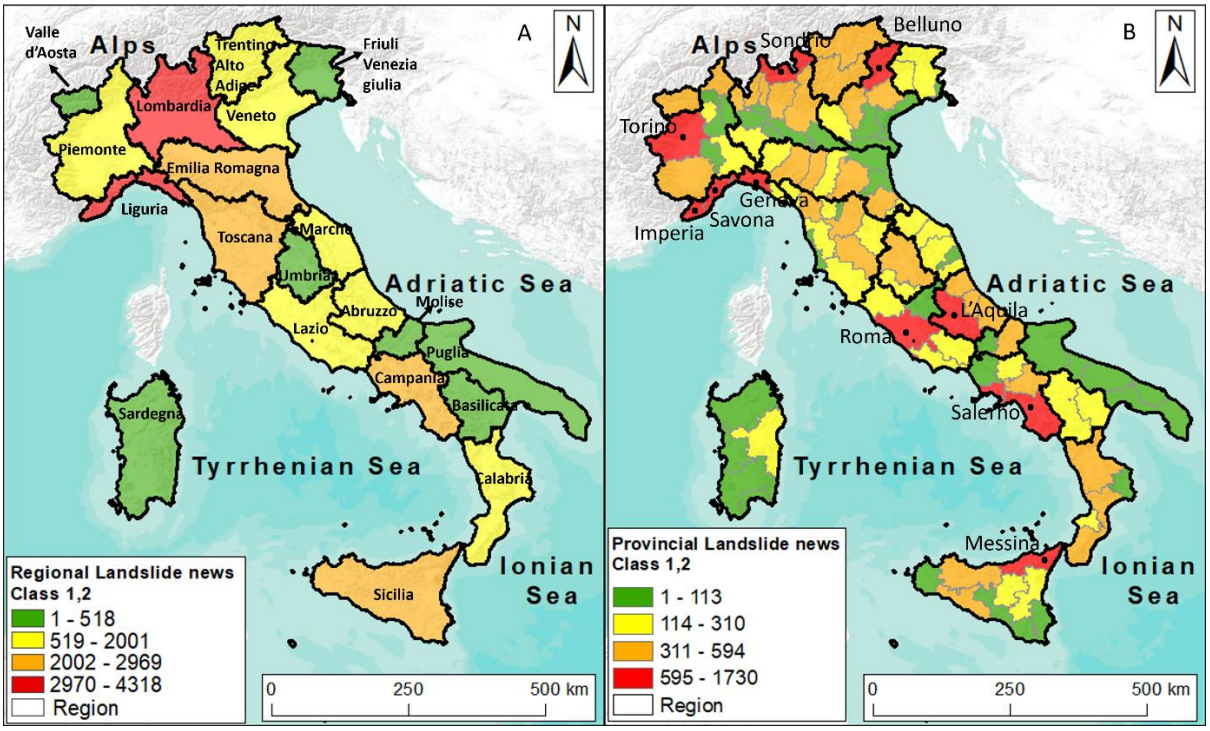

Figure 21: Spatial distribution of landslide news: **A**) Regional and in **B**) Province aggregation with overall news (classes 1, 2). Genova is the province most affected by landslides, followed by Salerno, Messina, Savona and Sondrio. The Puglia region and the provinces along the northeast coast show a lower number of landslide events. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

To analyse the landslide event distribution in detail, the counting of "Landslide events" and the sum of published news for each event were considered.

Figure 22 shows the distribution of only the Class 1 news (referring to recent "Landslide events") at the regional scale. Liguria is the region with the highest number of both articles and "Landslide events". Lombardia is the second region, regarding the number of "Landslide events" but with a lower number of articles. Toscana and Sicilia are the second and third regions, respectively, in terms of published articles. Valle d'Aosta, Friuli Venezia Giulia, Molise and Basilicata exhibit inferior media impact in agreement with the low trend of the "Landslide event".

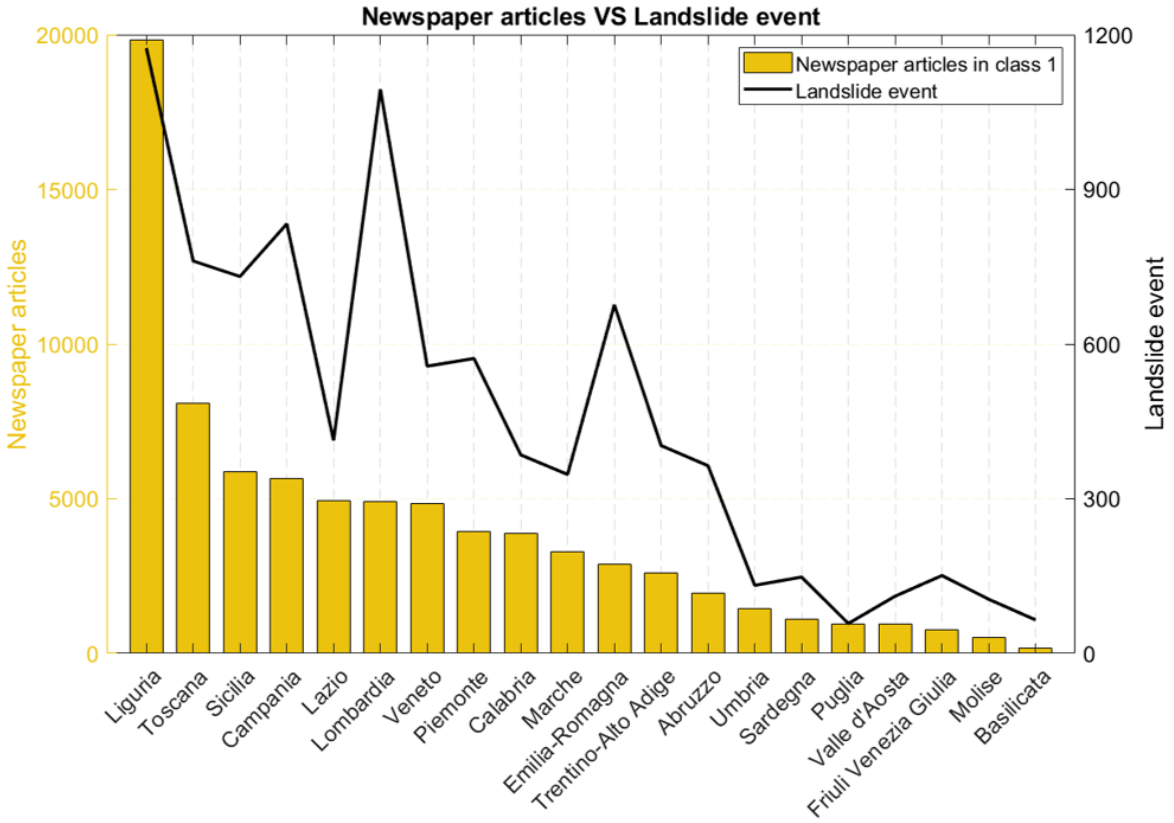

Figure 22: Regional distribution with a comparison between the number of published articles in Class 1 (media impact) and of "Landslide events". The panel was generated using MATLAB R2021b.

The number of days with at least 1 reported landslide event (landslide day) is higher in the northern regions than in the southern ones, except for Sicilia, the southernmost region, where a high number of landslide days is present (Figure 23A). Overall, 5 regions out of 20 had at least 450 days with landslide events in the analysed period. Lombardia, Liguria, Campania, Sicilia and Toscana are the regions with the highest number of days characterized by landslides. In particular, 677 days with landslides were identified in Lombardia, 572 in Liguria, 545 in Campania, 475 in Sicilia and 451 in Toscana (Figure 23A). The Puglia region has the lowest number of landslide days; in this region, 72 landslide events, distributed over 49 days, are present.

On a more detailed scale (Figure 23B), 4 provinces out of 107 have a high number of days with landslide events (180-301), while the average value is 23 days with landslides every year. For example, Genova Province is characterized by 915 landslide events, reported in 12.942 articles, distributed over 301 days. The provinces that have fewer days with at least one landslide event are located along the northeast coast, such as Venezia, Rovigo, Ferrara and Ravenna.

In general, the results show that Liguria, Lombardia, Campania, Toscana and Sicilia are the regions with the highest number of both "landslide events" and "landslide days".

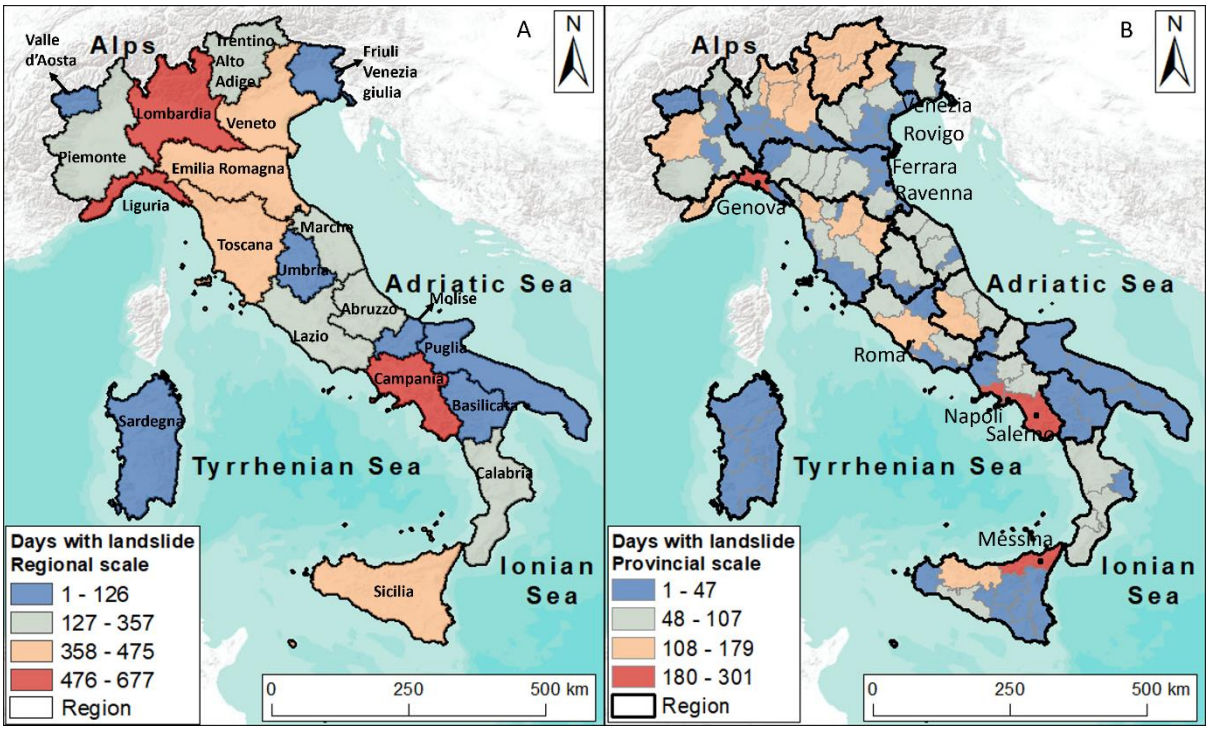

Figure 23: Spatial distribution of days with reported landslides. **A** Regional distribution, and **B** Provincial distribution. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

Temporal distribution

From a temporal point of view, the year with the highest number of landslide events showed a very sharp increase from 2017 (1243 events) to 2019 (2901 events) (Figure 24A), while the number of landslide-related articles was 2014 (Figure 24B).

Once a general overview of the spatial and temporal distribution of news has been accomplished, a more detailed analysis of only the Class 1 news has been carried out.

Figure 24C displays a monthly distribution of the landslide events identified by the Class 1 data; it shows that November, March and February are the months more involved with landslides.

November, indeed, for 10 years reported 2093 landslide events with 20142 published articles (multiple articles can refer to the same landslide event, as described in the previous section). July, June and September are the months with fewer events. For instance, in July, 597 “Landslide events” were reported by newspapers.

Class 1 news has been further analysed to identify the number of days with at least 1 landslide reported. The annual distribution (Figure 24D) follows a gradual increase of days with at least 1 landside from 2015 to 2019; in this period, 8103 landslide events have been collected, distributed over 1378 days, with an average of almost 5 landslides each day, while from 2010 to 2014, 5172 landslide events, distributed over 1236 days, were reported.

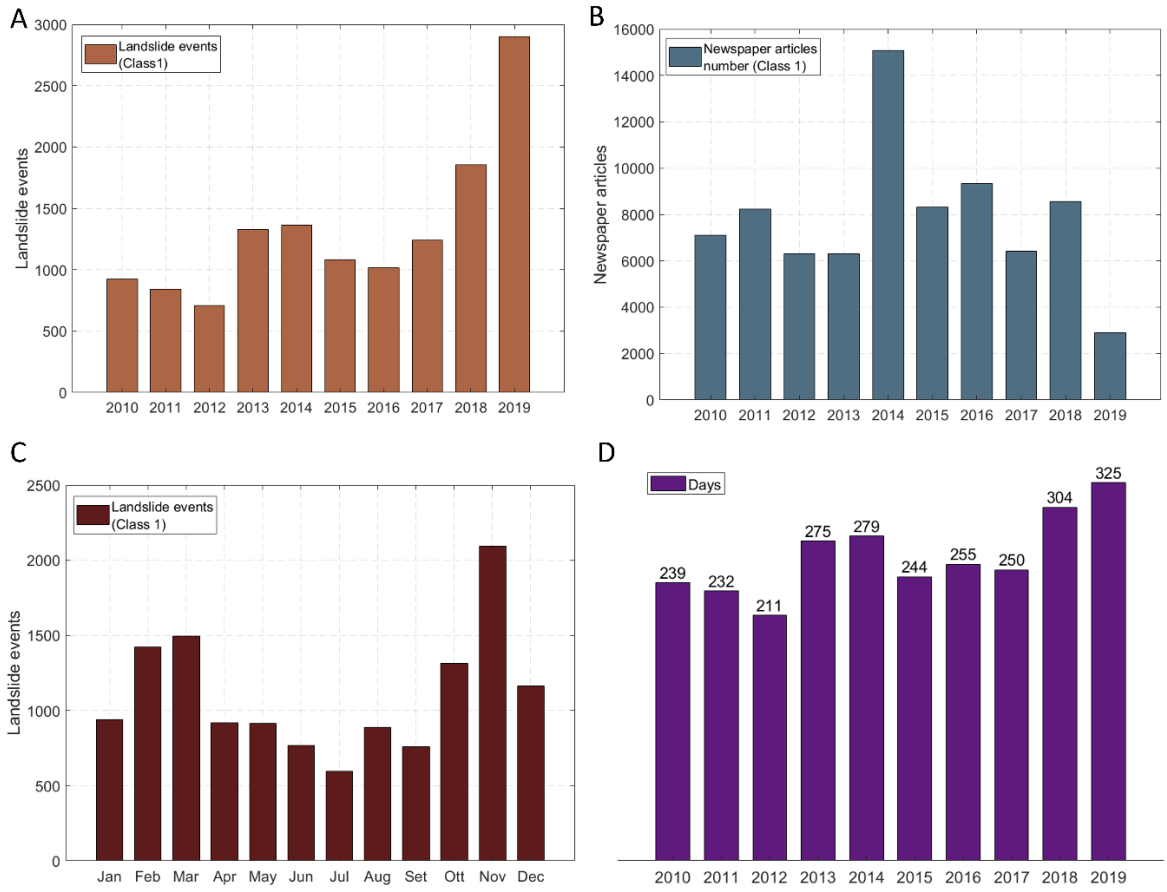

Figure 24: Temporal distribution of Class 1 news. **A** "Landslide events" annual distribution; **B** "Newspaper article" annual distribution; **C** monthly distribution of "landslides events"; **D** The number of days with at least 1 landslide reported from 2010 to 2019. Panels were generated using MATLAB R2021b.

4.1.1.1 Correlation with traditional sensors

Several organizations create reports or datasets for describing many different aspects of natural disasters. In this work, four 10-year databases (2010-2019) were analysed for landslide events in Italy. The analysis was carried out to obtain information and determine the spatial, regional and WHZ scales and the temporal correlations of the available data. Overall, the analysed data are divided as follows: 31.878 "Landslide news" with 174.616 "Newspaper articles", 2040 rain gauges with 35.299 rainfall events, 198 data from Polaris and 1539 data from ReNDiS. All the datasets cover the period 2010-2019, except the Polaris dataset, since it starts from 2011.

Spatial distribution

The “Landslide news” cannot identify the exact location of a landslide, and the maximum spatial result that can be obtained is the municipality. For this reason, the data have been grouped on regional and WHZ bases to outline the areas with a higher number of published articles. The number of ”Landslide news” was used as a proxy to identify those areas more affected by landslides in the observed period, hence the most hazardous areas, while the number of ”Newspaper articles” was used as an estimator for landslide intensity, since severe landslides can lead to more catastrophic effects and have a higher media echo (hence, a higher number of referencing articles)

The regions most affected by “Landslide events” were mainly in the northern portion of the country. Liguria and Lombardia have been the regions with the highest amount of “Landslide news” (Figure 25A). A similar spatial distribution was achieved considering the rainfall data (Figure 25B), relevant rainfall distribution (Figure 25C) and IDMs (Figure 25D). Indeed, the Northern Regions registered the highest mean annual rainfall and relevant rainfalls with respect to the Central or Southern Regions, except for Campania and Calabria. Friuli Venezia Giulia, Lombardia and Piemonte were the rainiest regions of Italy. The distribution of IDMs was more even across the country but it also reveals that the northern area experienced a higher number of IDMs. For example, Trentino Alto Adige, Lombardia and Liguria were the three regions with the highest IDM numbers, 31, 30 and 27, respectively. Furthermore, the first two regions have been the regions most involved for 10 years (Figure 25E). No region showed 0 IDM after a landslide. Basilicata, Molise and Puglia showed only one IDM in agreement with the low values of the landslide events and rainfall data. Nonetheless, Figure 25F shows the earmarked funds for each region, in which the Campania, Sicilia and Puglia regions stand out for the highest funds allocation of the country. Other Southern Regions, such as Basilicata and Calabria, have shown strong investments following landslide events.

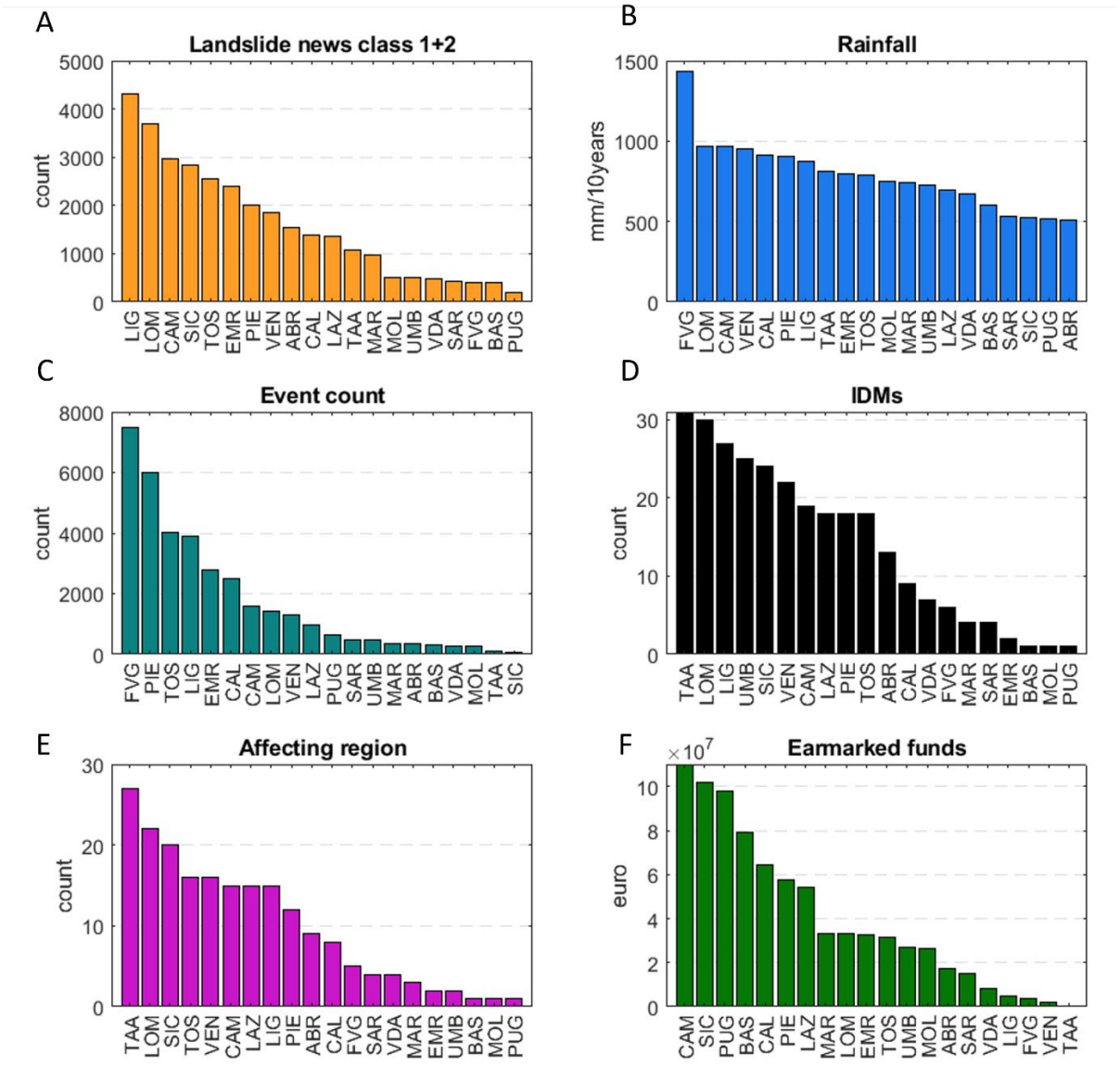

Figure 25: Spatial distribution with different information: **A** "Landslide news" and "Newspaper articles" from Social Media; **B** Rainfall data with mm/y and in **C** event counts; **D** Polaris with IDEMs (injured, death, evacuated and missing) number; **E** involved Region; **F** earmarked funds for the soil protection (euro – ReNDiS with a focus for better vision). **ABR**: Abruzzo, **BAS**: Basilicata, **CAL**: Calabria, **CAM**: Campania, **EMR**: Emilia-Romagna, **FVG**: Friuli-Venezia Giulia, **LAZ**: Lazio, **LIG**: Liguria, **LOM**: Lombardia, **MAR**: Marche, **MOL**: Molise, **PIE**: Piemonte, **PUG**: Puglia, **SAR**: Sardegna, **SIC**: Sicilia, **TOS**: Toscana, **TAA**: Trentino-Alto Adige, **UMB**: Umbria, **VDA**: Valle d'Aosta, **VEN**: Veneto. The arrow indicates the increasing direction of the allocated funds. Panels were generated using MATLAB R2021b.

In general, the distribution of "Landslide news" (Figure 26A) and "Newspaper articles" (Figure 26B) agreed with the number of relevant rainfalls (Figure 26C-D) and IDMs (Figure 26E) but in some cases, it was in contrast with earmarked funds (Figure 26F). In fact, the earmarked funds for soil protection outlined an inverse distribution, showing more investments in southern Italy than in northern Italy.

Valle d'Aosta, Molise and Abruzzo have been the regions that showed the best coherence between the variables. In this sense, the morphology of the territory, the climatic conditions, the size of the region and the density of the people and buildings at risk can bias the distribution of landslide events. In conclusion, Valle d'Aosta, Piemonte, Liguria, Lombardia, Veneto, Emilia-Romagna, Toscana, Marche, Abruzzo, Lazio, Molise, Campania and Sardegna have been the regions that showed a higher coherence between the datasets. Conversely, Friuli Venezia Giulia, Trentino Alto Adige, Umbria, Puglia, Basilicata, Calabria and Sicilia have been the regions with lower coherence.

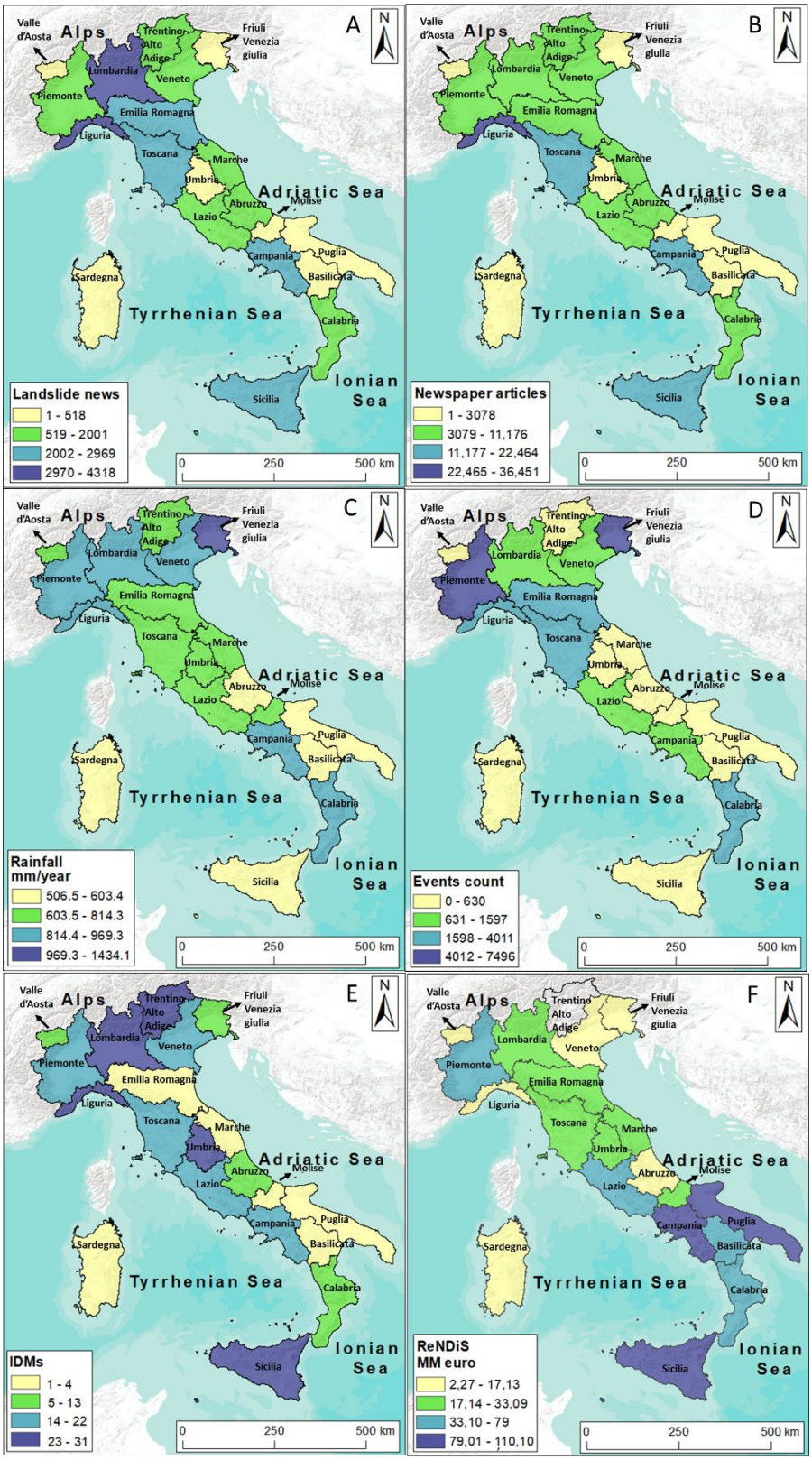

Figure 26: A Regional aggregation with "Landslide news", estimating landslide hazard; **B** Regional aggregation with mediatic impact, estimating the landslide intensity; **C** Regional distribution with rainfall data for 10 years (mm/10 years); **D** Regional aggregation of relevant rainfall event counts. Events display relevant differences in spatial distribution. **E** Regional aggregation with IDMs from the Polaris dataset. The period covers only 9 years, from 2011 to 2019. **F** Regional aggregation with earmarked funds for soil protection from the ReNDiS (euro/10 years) for 10 years, considering landslide events. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

“Newspaper articles”, rainfall distribution and earmarked funds have been normalized on the basis of the annual maximum to analyse and compare the variation of their values over time for each region. Figure 27 shows the trend of each variable for each region over 10 years. Regarding the Northern Regions, Trentino Alto Adige, Valle d’Aosta and Veneto did not show variations in earmarked fund distributions. Friuli Venezia Giulia, Lombardia and Piemonte reveal the same distribution of earmarked money during 2015 (Figure 27FVG and LOM). The Piemonte Region showed a sharp increase in investment in soil protection in 2018. The regions Lombardia, Piemonte, Trentino alto Adige, Valle d’Aosta and Veneto showed the same trend between “Newspaper articles” and rainfall distribution; the lowest rainfall corresponded to the lowest media impact on landslide events (Figure 27LOM, PIE, TAA, VDA and VEN). In contrast, for Friuli Venezia Giulia, it was possible to recognize high values of rainfall but very low values of the media impact of the landslide events (Figure 27FVG).

Abruzzo, Emilia Romagna, Lazio, Marche, Sardegna, Toscana and Umbria have been the regions with a good correlation with the annual distribution of the variables. In fact, the “Newspaper articles” number increases or decreases as a consequence of rainfall data, and the earmarked funds increase in the same year or in the following years (e.g., Abruzzo in 2012 and 2015 (Figure 27ABR), Emilia Romagna in 2010, 2013, 2015, 2016 and 2017 (Figure 27EMR), Lazio in 2010, 2014, 2015 and 2018 (Figure 27LAZ), Marche in 2013, 2014, 2015 and 2017 (Figure 27MAR), Sardegna in 2013 and 2015 (Figure 27SAR), Toscana in 2010, 2013, 2015 and 2016 (Figure 27TOS), Umbria in 2013, 2015 and 2017 (Figure 27UMB)).

Basilicata, Calabria, Campania, Molise, Puglia and Sicilia showed a good correlation between variables for 10 years. Basilicata and Puglia evinced similar trends for each variable. All of them revealed low values of “Newspaper articles” but a good correlation between rainfall data and earmarked funds. In fact, high values of rainfall were measured during 2013, 2015 and 2018 and the distribution of earmarked funds in the same year (e.g., Puglia in 2015, Figure 27PUG) or in the next year (Basilicata in 2016 and 2019 and Puglia in 2017 in Figure 27BAS and PUG). Calabria, Campania, Molise and Sicilia presented similar distributions for each variable. For example, during 2010, “Newspaper articles”, rainfall data and earmarked funds revealed high values in each region except for the Molise; during 2013, the increase in “Newspaper articles” coincided with the increase in rainfall data but corresponded to low values of earmarked funds (fund increments started from 2014 in Sicilia and only in 2015 in the other regions). 2015 and 2018 showed the most coherence; in fact, all the southern regions featured high values of rainfall data and “Newspaper articles”.

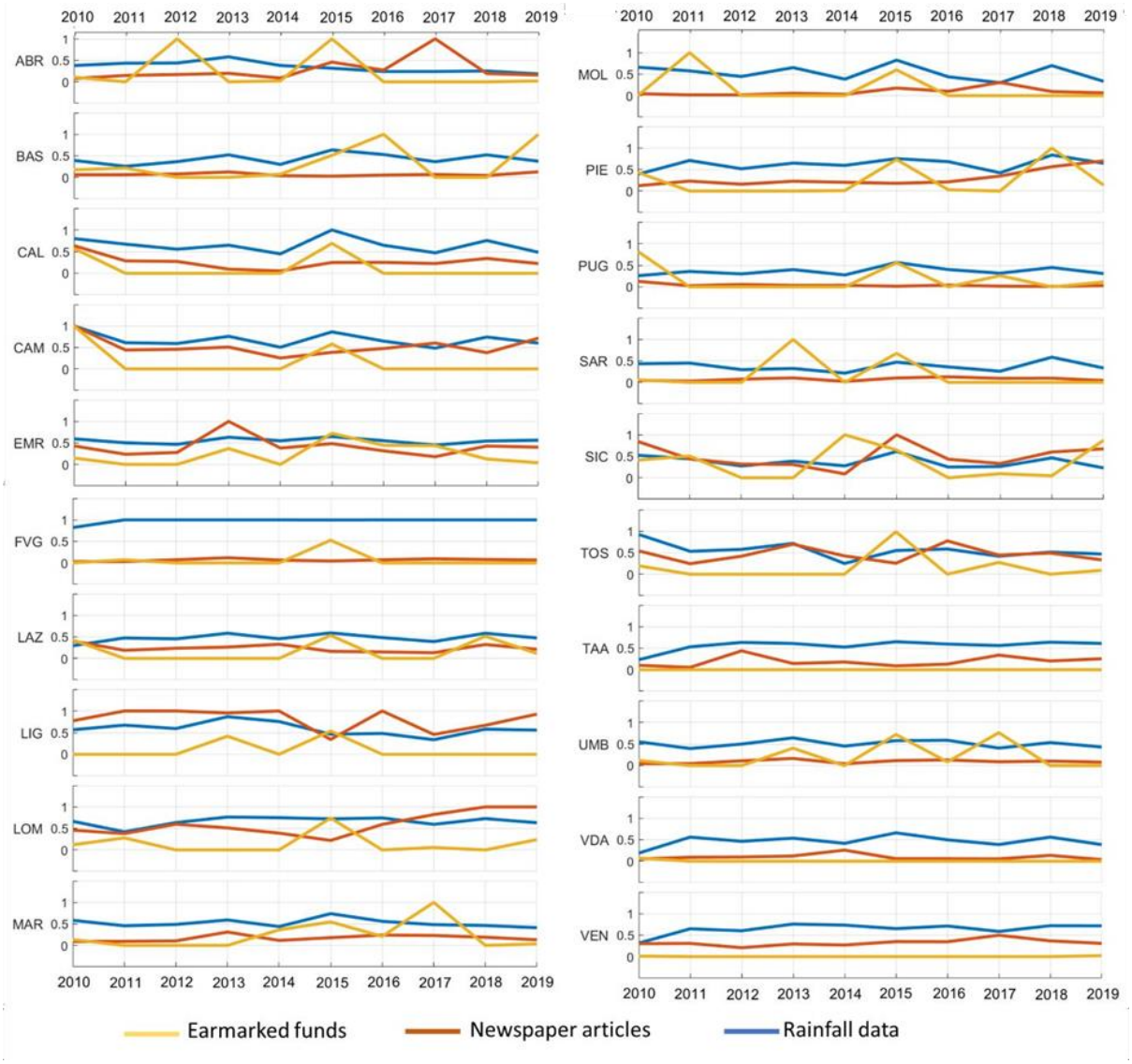

Figure 27: "Newspaper articles", rainfall distribution and earmarked funds have been normalized and correlated for each region for 10 years. **ABR**: Abruzzo, **BAS**: Basilicata, **CAL**: Calabria, **CAM**: Campania, **EMR**: Emilia-Romagna, **FVG**: Friuli-Venezia Giulia, **LAZ**: Lazio, **LIG**: Liguria, **LOM**: Lombardia, **MAR**: Marche, **MOL**: Molise, **PIE**: Piemonte, **PUG**: Puglia, **SAR**: Sardegna, **SIC**: Sicilia, **TOS**: Toscana, **TAA**: Trentino-Alto Adige, **UMB**: Umbria, **VDA**: Valle d'Aosta, **VEN**: Veneto. Panels were generated using MATLAB R2021b.

Correlation with hazard maps

To validate the quality of the results, mainly of the spatial distribution of "Landslide news", a comparison with existing datasets about landslides has been made. The landslide hazard map of Italy (Trigila et al., 2018), the map of populations living in landslide-risk areas (Trigila et al., 2018) and the map of building at risk (Trigila et al., 2018) have been used. These 2 maps were processed to extract the percentage with respect to the total Italian territory (300.000 km^2) and the number of inhabitants (59 million). The percentage of area of each region affected by landslide hazards (Figure 28B) and the percentage of the population of each region living in zones affected by landslide risk (Figure 28C) were calculated. This operation was needed to account for the differences in size and population of the

different regions, which can vary greatly. Some large regions (e.g., Lombardia and Emilia-Romagna) are characterized by a high percentage of landslide hazards but a low percentage of people at risk. This trend is because Lombardia and Emilia Romagna feature wide plain areas with significant urbanization. The comparison between the three maps in Figure 28 shows good agreement between the distributions of “Landslide news”, landslide hazard or people at risk, but some anomalies can be identified. For instance, Valle d’Aosta shows a lower number of “Landslide news” but a very high portion of the territory is subject to landslide hazard. Another example, Sicilia, which has an important amount of news and a low percentage of its territory subject to landslide hazards, but a significant number of people live in hazard areas. The distribution of building at risk (Figure 28D) shows import values of percentage in P3 and P4 in Campania and in Toscana with corresponding with all variables. Lombardia shows trend inverse of percentage of building at risk respect to “Landslide news” and landslide hazard. Puglia, Basilicata and Calabria present opposite trend with low values of “Landslide news”, but with significant percentage of building at risk.

Figure 28: Comparison between the distribution of landslide news (classes 1 and 2, Panel **A**), landslide hazard (Panel **B**), people at risk (Panel **C**) and building at risk (Panel **D**). The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

The number of "Landslide news" items has been correlated with the aforementioned percentages to better verify the existence of a correspondence between "Landslide news", landslide hazard and population at risk. Figure 29A shows a general correlation between the number of news items (Class 1+2) and the areas affected by landslide hazards for each region (with nonparametric correlation of Kendal= 0,24 , Spearman= 0,32 and R= 0,29), and Figure 29B shows the population living at risk (with nonparametric correlation of Kendal= 0,51 , Spearman= 0,67 and R= 0,55). The distribution of the data shows some anomalies that are due to the morphology of the territory and the involvement of habituated areas. Toscana, Emilia-Romagna, Campania, Lombardia, Sicilia and Liguria are characterized

by significant values of landslide hazard similar to the published articles. The correlation between "Landslide news" and the percentage of people at risk is even more marked. This relationship can be related to the fact that the greater the urbanisation, the greater the number of articles published on the landslide event. (Campania, Toscana, Sicilia and Liguria). The number of "Landslide news" has been furthermore correlated with the percentages of building at risk (Figure 29C). The correlation presents a general correlation between the number of news (Class 1+2) and building in hazard areas for each region. In this case nonparametric correlation have been calculate and results are respectively with Kendal almost 0,42 , Spearman of 0,60 and R of 0,58. Table 13 lists for each correlation between "Landslide news" and the percentage of landslide hazard area, people at risk and building at risk, non parametric coefficients Kendall, Spearman and Pearson.

Landslide news (Class1+2) VS	**Kendal-K**	**Spearman-S**	**Pearson-R**
Percentage of landslide hazard area	0,24	0,32	0,29
Percentage of people at risk	0,51	0,67	0,55
Percentage of building at risk	0,42	0,60	0,58

Table 13: Non paremetric coefficients, Kendal, Spearman and Pearson, for each correlation have showed.

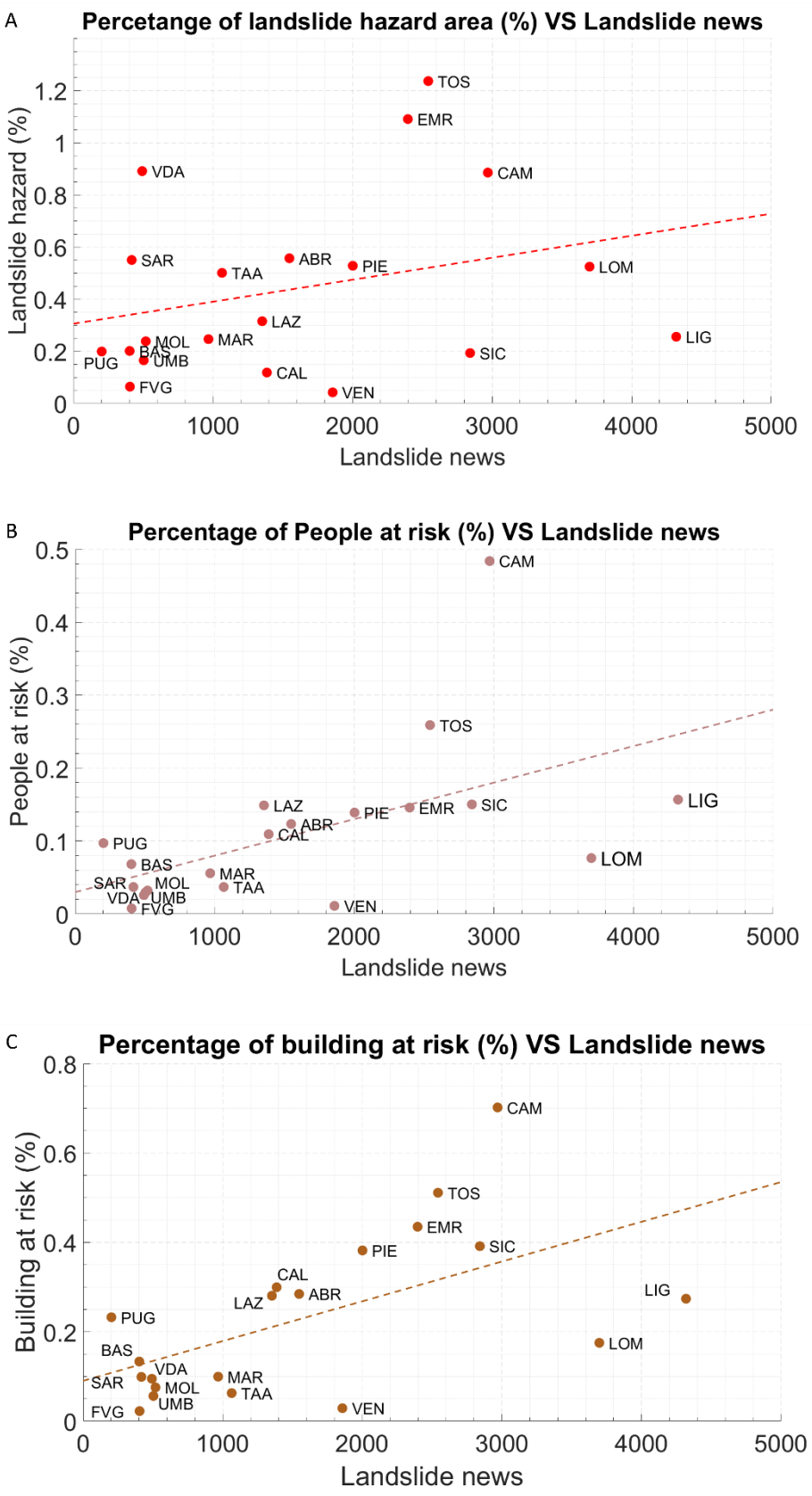

Figure 29: Correlation between the number of "Landslide news" (Class 1+2 and the percentage of the landslide hazard area (**A**) and the percentage of people at risk (**B**). **ABR**: Abruzzo, **BAS**: Basilicata, **CAL**: Calabria, **CAM**: Campania, **EMR**: Emilia-Romagna, **FVG**: Friuli-Venezia Giulia, **LAZ**: Lazio, **LIG**: Liguria, **LOM**: Lombardia, **MAR**: Marche, **MOL**: Molise, **PIE**: Piemonte, **PUG**: Puglia, **SAR**: Sardegna, **SIC**: Sicilia, **TOS**: Toscana, **TAA**: Trentino-Alto Adige, **UMB**: Umbria, **VDA**: Valle d'Aosta, **VEN**: Veneto. Panels were generated using MATLAB R2021b.

Subsequently, the correlation between the earmarked funds and the percentages of hazardous areas and buildings at risk of each region was analysed (Figure 30). Since the extent of the Italian regions varies greatly, the percentages of hazardous areas and buildings at risk have been scaled with respect to the area of the Italian territory and to the total number of buildings.

The analysis has been carried out considering the distribution of funds at the regional scale. The distribution of funds allowed outlining which regions showed a higher percentage of both territories subject to landslide hazard and buildings at risk compared with the funds allocated for soil protection. For these analyses, the Italian regions were split according to their geographical distribution (north, central and south).

In some cases, low values of allocated funds were in agreement with the percentages of hazardous areas and buildings at risk, such as for Veneto and Friuli Venezia Giulia. These regions have been mainly involved in other natural events, such as floods and earthquakes. The trend was the opposite for the Valle d'Aosta and Liguria regions (Figure 30A-B in blue), both having a high portion of their territory subject to landslide hazards but few earmarked funds. Instead, Lombardia and Piemonte (both partially plain areas) showed high values of funds, with intermediate values of the percentage of landslide hazard areas and buildings at risk. This may be linked to the urbanisation and the population density spread in the floodplain in southern Lombardia and Piemonte, while landslides are concentrated in the northern part of the region along the Alpine arc, where they caused many fatalities.

In Central Italy, a more homogeneous distribution can be recognized, with values ranging from 15 to 54 million €. The percentage of hazardous areas ranged from 0,16% in Umbria to 1,2% in Toscana, while the percentage of buildings at risk was 0,05% and 0,51% in the same regions (Figure 30A in green). Emilia Romagna and Abruzzo follow the Toscana Region, showing high values of landslide hazard (approximately 1% for the first and approximately 0,55% for the second). In these regions, the high percentage of hazard entailed a high percentage of buildings at risk, 0,43% and 0,28%, respectively. The high percentages of the Emilia Romagna Region were in agreement with the higher values of allocated funds 32.6 million (Figure 30A in green).

Southern Italy has been, in general, the portion of Italy with the most earmarked funds (Figure 30 in light blue). The earmarked funds were in agreement with the highest percentage of landslide hazards in the Campania Region. In general, the percentage of landslide hazard areas varied from a minimum of almost 0,11% in Calabria to a maximum of 0,88% in Campania. Campania, Molise and Basilicata have been the regions with the highest percentage of landslide hazard areas between the southern regions. Furthermore, the allocated funds concurred with the percentage of buildings at risk in all the southern regions. The Calabria, Sicilia and Campania regions showed the highest percentages of buildings at risk, 0,29%, 0,39% and 0,70%, respectively.

This analysis revealed that the southern regions had, in the observed periods, more funds for soil defence, even if the number of news items related to landslides and the percentages of territory subject to landslides and of buildings at risk were sensibly lower than those of other parts of Italy.

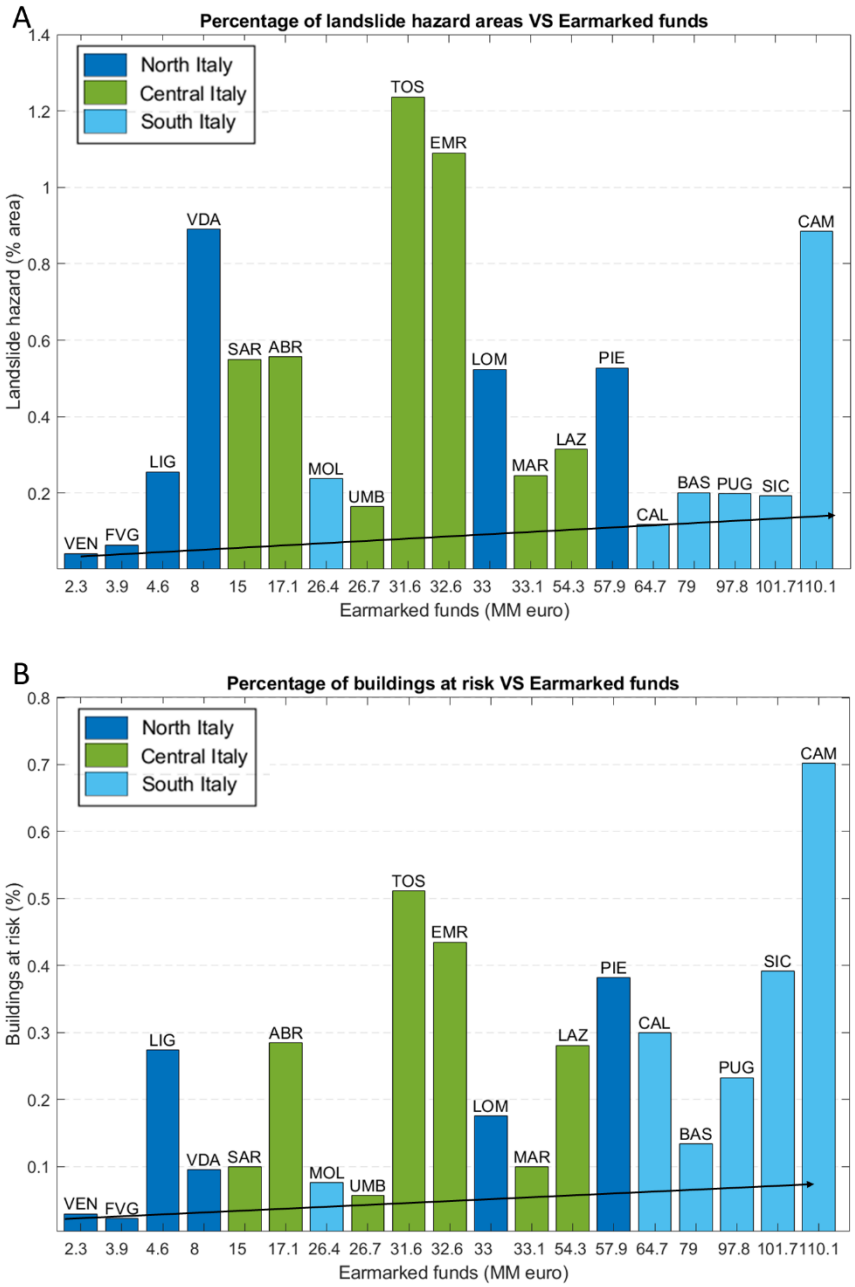

Figure 30: Validate data between earmarked funds in **A** with the percentage of landslide hazard areas for each region; in **B,** the percentage of buildings at risk for each region. Each percentage was provided by ISPRA and normalized on the basis of regional size. **ABR**: Abruzzo, **BAS**: Basilicata, **CAL**: Calabria, **CAM**: Campania, **EMR**: Emilia-Romagna, **FVG**: Friuli-Venezia Giulia, **LAZ**: Lazio, **LIG**: Liguria, **LOM**: Lombardia, **MAR**: Marche, **MOL**: Molise, **PIE**: Piemonte, **PUG**: Puglia, **SAR**: Sardegna, **SIC**: Sicilia, **TOS**: Toscana, **TAA**: Trentino-Alto Adige, **UMB**: Umbria, **VDA**: Valle d'Aosta, **VEN**: Veneto. The arrow indicates the increasing direction of the allocated funds. Panels were generated using MATLAB R2021b.

Warning hydrological zones (WHZs)

For civil protection purposes, knowledge of the rain distribution is fundamental when preparing national weather bulletins. The combination of rain distribution data with "Landslide news" and "Newspaper articles", at the WHZ scale, can provide helpful outcomes for those involved in landslide management and prevention. “Landslide news” and “Newspaper articles” were mainly located in the northern part of Italy and along the Apennines (Figure 31A-B). In general, the number of “Landslide news” articles showed a direct correlation with the number of “Newspaper articles” for each WHZ. An exception can be recognized for one WHZ in the Puglia Region (area north), where a low number of “Landslide news” was associated with a relatively higher number of ”Newspaper articles”, meaning that landslide events had a high media echo in this area.

The areas less involved with landslides were located along the northeast coast and in Puglia, and a similar distribution was recognized by Del Soldato et al. (2021) considering the rainfall frequency. Indeed, North Italy showed a very good correspondence between the variable “Newspaper articles” and rainfall data, with an emphasis on the Liguria Region, north-western portion of the Alps (Valle d’Aosta and Piemonte) and the northeast (Trentino Alto Adige, Veneto and in part of Friuli Venezia Giulia).

Central Italy was the portion where the main differences were recognized between the different intensities of rainfall (Figure 31C-D-E). In fact, the rainfall with “High” intensity has shown the most distribution than “Very high” and “Heavy rain”, which has concentrated the most in Liguria and in the most north Toscana Region. A correspondence has been highlighted between “Newspaper articles” and the intensity classes “High”, “Very high”, “Heavy rain”, as well as between “Newspaper articles” and “events count” (Figure 31F) in the northern WHZs of the Toscana and Emilia Romagna Regions and in Central Italy with Marche, Abruzzo and Lazio.

In southern Italy, the highest values of “Landslide news”, “Newspaper articles” and rainfall data have been located along the Ionian seacoast, in the WHZs of the Calabria Region and in one WHZ of the Basilicata Region (south-eastern area). The Puglia Region did not show high values of “Landslide news”; however, in one area, a high mediatic impact (“Newspaper articles”) was revealed, even when associated with low frequencies of relevant rainfall events.

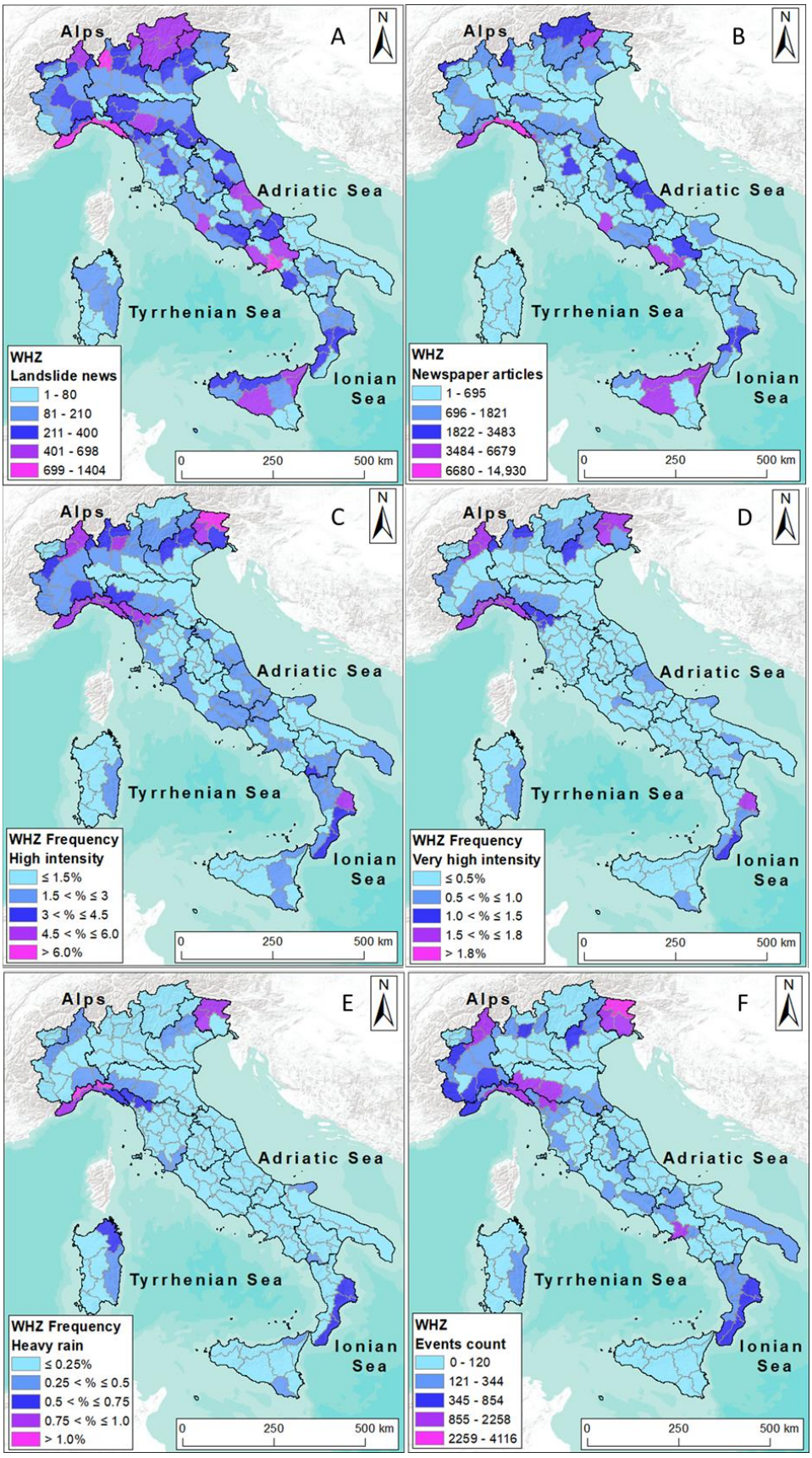

Figure 31: Italy is divided into 158 Warning hydrological zones (WHZs). In **A** the distribution of "Landslide news"; **B** the distribution of media impact with "Newspaper articles"; **C-D-E** the rain frequency "High", "Very high" and "Heavy rain"; **F** the events sum of three classes or events count. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

Figure 32 shows the correlation between “Newspaper articles” and several rainfall intensities (Figure 32A-B-C) and the event count (Figure 32D) for each WHZ. Although low values of media impact (as “Newspaper articles”) sometimes corresponded to low values of rainfall intensity and vice versa, it was not possible to outline a clear correlation.

Since the data do not follow a Gaussian distribution, two nonparametric correlation indices were used to verify the rate of correlation between the analysed parameters. The Kendall’s and Spearman rank correlation coefficients resulted in low values, as reported in Table 14, confirming the presence of a slight correlation between the parameters.

	Kendall	**Spearman**
Newspaper articles – High Intensity rainfall	0,15	0,22
Newspaper articles – Very high Intensity rainfall	0,09	0,14
Newspaper articles – Heavy rain	0,13	0,19
Newspaper articles – Events count	0,20	0,29

Table 14: Kendall and Spearman coefficient rank for each correlation between variables.

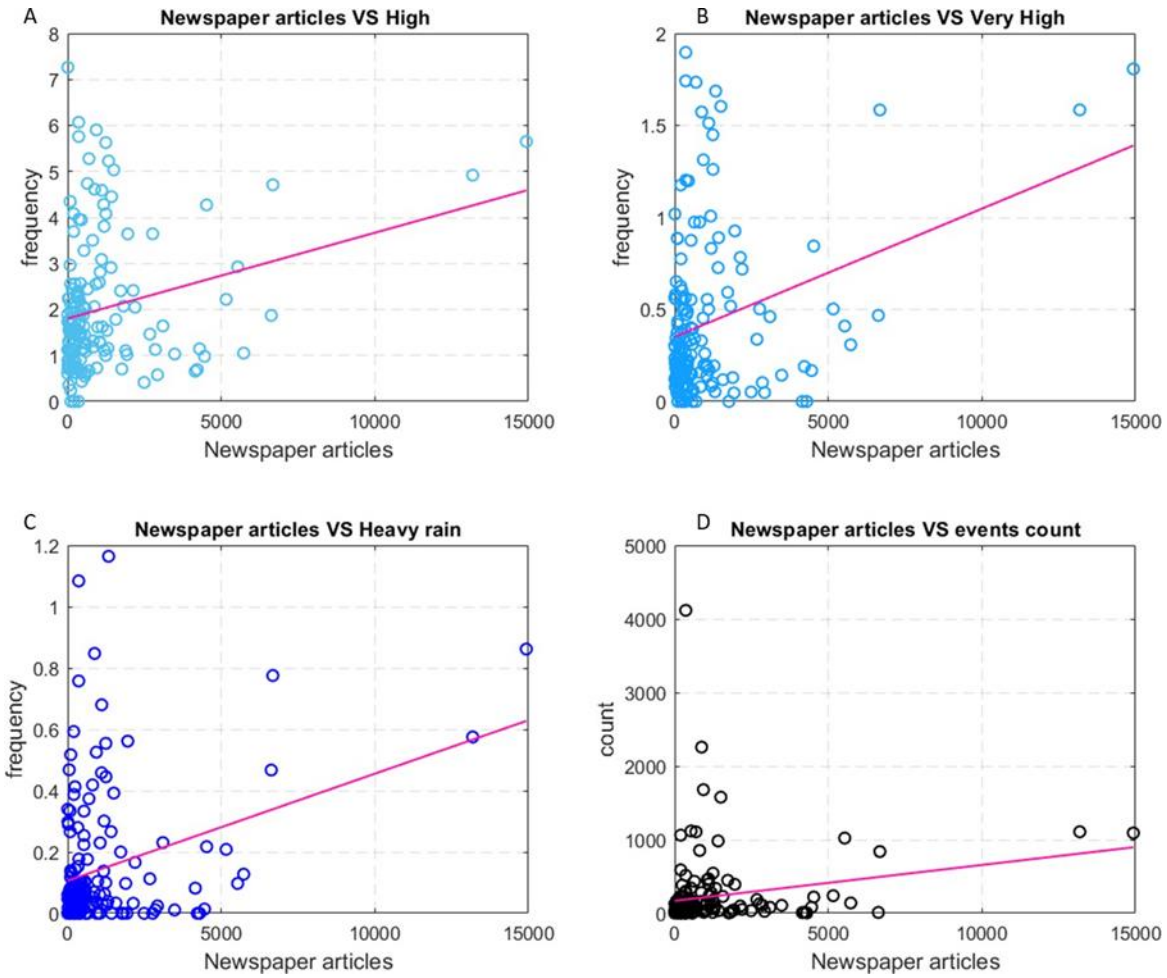

Figure 32: Distribution of "Newspaper articles" with different frequencies of rainfall intensity for each WHZ. In **A** "Newspaper articles" and "High" intensity; **B** "Newspaper articles" and "Very high" intensity; **C** "Newspaper articles" and "Heavy rain"; **D** the correlation between "Newspaper articles" and the events count. Panels were generated using MATLAB R2021b.

Temporal distribution

From a temporal point of view, each Italian region experienced some landslides in the investigated period, with approximately 1477 IDEMs per year.

"Landslide news" showed an increasing trend from 2010 to 2014, which repeats in the period from 2015-2019. At the end of the first quinquennium, 2014 featured the highest number of "Landslide news" (2988), with a mean rainfall of 1007,6 mm/year (Figure 33A-B). 2014, with 4706 rain events, was also the year with the highest number of relevant rainfall events (Figure 33C). In this year, 19 regions out of 20 were involved in landslides with 3406 IDEMs (Figure 33D-E) as consequences. In contrast, 2014 was the second year with less earmarked funds for soil protection, with almost 8 million euros, second only to 2012, with only 227 thousand euros (Figure 33F).

In the second quinquennium, the year with the highest number of "Landslide news", with 6775 data, was 2019. Although the rainfall data were constant with respect to the previous year, 2019 was also

the second year with a higher event count, with 4529 relevant rain events (Figure 33A-B-C). In the same year, 16 regions were affected by landslides with 2775 IDEMs (Figure 33D-E). 2019, in these five years, showed the highest values of earmarked funds, with almost 109 million euros (Figure 33F).

In general, the “Landslide news” showed an increase from 12.083 news items in the period 2010-2014 to 19.795 in 2015-2019, while for the rainfall events, the trend was the inverse. The count of relevant rainfall events decreased from 18.858 for 2010-2014 to 16.441 for 2015-2019, passing from an average of 3771 events/year to 3288 events/year, and the rainfall data passed from 4451 mm/5 years to 3910 mm/5 years. The same decrease distribution was observed in “Newspaper articles”, IDEMs and reported expenses between the periods 2010-2014 and 2015-2019. “Newspaper articles” featured 91.439 articles for the first quinquennium to 83.177 in the second, while the IDEMs distribution showed a small decrease from 7402 to 7372.

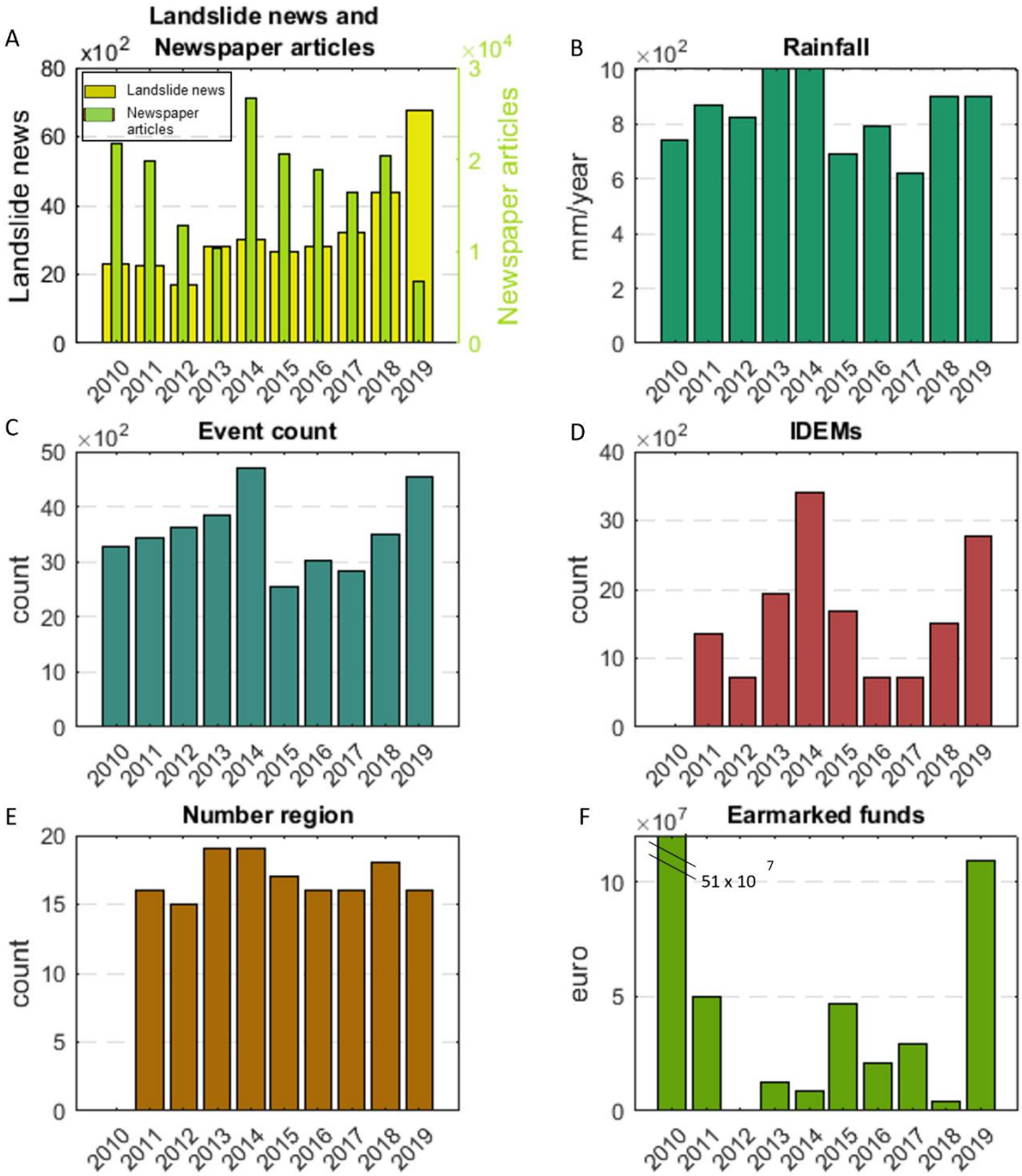

Figure 33: Temporal distribution with different information: **A** "Landslide news" and "Newspaper articles" from Social Media; **B** Rainfall data with mm/y and in **C** event counts; **D** Polaris with IDEMs (injured, death, evacuated and missing) number; **E** involved region; **F** earmarked funds for the soil protection (euro – ReNDiS with a focus for better vision). Panels were generated using MATLAB R2021b.

4.1.1.2 Text analysis and word distribution

From 2010 to 2019, 32.525 news items were gathered by the data mining algorithm.

The dataset provides the headline as the only textual source. In the beginning, the text was preprocessed by removing all textual parameters lacking literary meaning within the sentences; this

included articles, punctuation, special characters, number of words with less than 2 letters, and low word frequency. Once the text was cleaned, two analyses were applied, a qualitative analysis with a word cloud (Figure 34A) and a quantitative analysis with a word frequency count (Figure 34B).

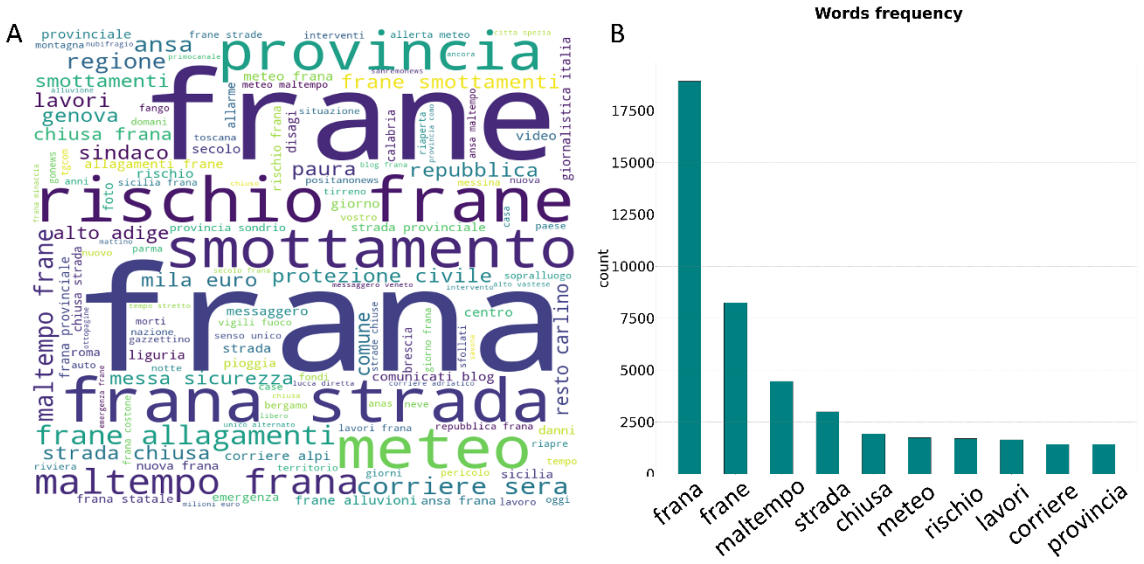

Figure 34: Qualitative representation using a word cloud in **A,** and in **B** a quantitative representation with the word frequency with the first ten words. Panels were generated using Python. Below the translation for each word; frana: landslide, frane: landslides, maltempo: bad weather, strada: road, chiusa: close, meteo: weather, rischio: risk, lavori: jobs, Corriere: newspaper name, provincial: province.

Among them, 13.275 news items had useful information about the geo-localization and the date of the landslide event; 1400 news items have been corrected, attributing a more appropriate localization based on the text to the article. According to the adopted classification criteria, the identified news has been classified as follows:

- Class 1: 13.275 news items (41%).
- Class 2: 18.603 news items (57%).
- Class 3: 647 news items (2%).

This classification allowed us to identify the “true news” (classes 1 and 2) and to reject the data that were not appropriate (Class 3), reducing the data to be processed. Approximately 41% of news reported information relative to recent landslides, and only a minimum percentage of the database is made up by wrong news (2%) (Figure 35A). A textual analysis was conducted to retrieve the frequency of words inside the headlines. In Figure 35B-C-D, the most frequent words of the headlines of the Class 1, 2 and 3 news are reported, respectively. The term “landslide” is present in all categories as the first word widely used; indeed, in Class 1, the word “landslide” is present 8021 times, 10.457 times in class 2 and 271 times in class 3.

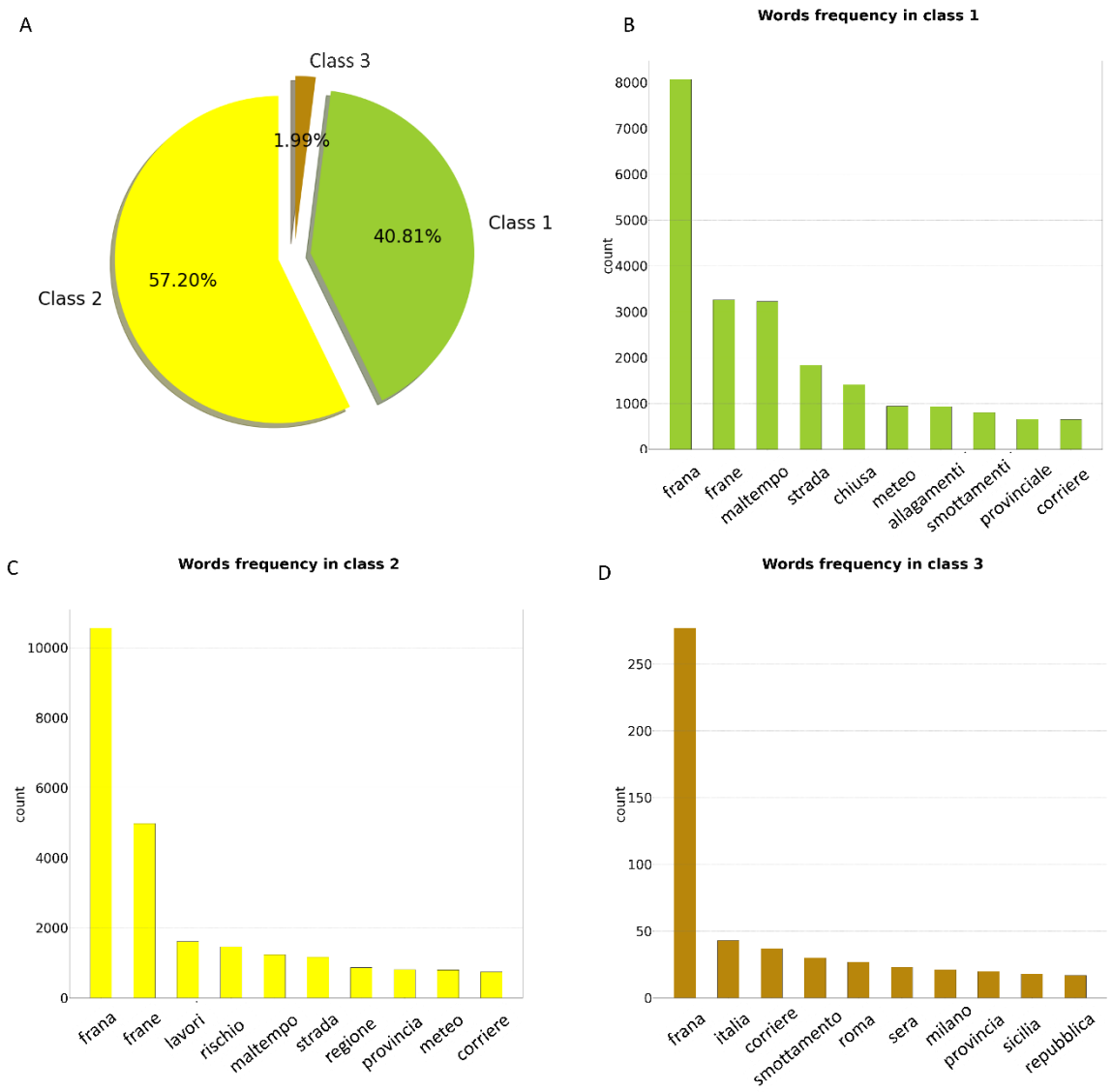

Figure 35: **A** Overall landside news classification. **B** Word frequency in the headlines inside Class 1, **C** Word frequency in the headlines inside Class 2, and **D** Word frequency in the headlines inside Class 3. Panels were generated using Python. Below the translation for each word; frana: landslide, frane: landslides, maltempo: bad weather, strada: road, chiusa: locked, meteo: weather, allagamenti: flooding, smottamenti: landslips, provinciale: provincial, rischio: risk, lavori: works, regione: region, Corriere: newspaper name, smottamento: landslip, provincial: province, Italia: Italy, roma: roma, sera: evening, milano: Milano, sicilia: Sicilia, repubblica: repubblic.

4.1.2 Flood news

Spatial distribution

The regions most involved are mainly in the Central area of the Country. Liguria and Toscana are the regions with the highest number of news (class 1 and 2) and therefore of "Newspaper articles" (articles referred to the same flood event are grouped into a single "Flood news"). For example, Liguria has 55.666 articles referring to 4977 "Flood news" (classes 1 and 2, Figure 36A), among them 11.157 articles refer to 490 recent "Flood events" (class 1) and in particular, Genova is the most affected province (Figure 36B).

Provinces showed a relevant number of news (Roma and L'Aquila) and they are mainly located along Tyrrhenian Sea coast and mains alluvial planes as in Emilia Romagna, Veneto and internal flat areas in Toscana, Piemonte and Campania regions (Figure 36B).

The Northern region (Valle d'Aosta, Trentino Alto Adige and Friuli Venezia Giulia) (Figure 36A) and the provinces along the South-East coast along Ionian Sea coast (Figure 36B) showed a lower number of flood news because they are mainly mountain or hill areas and fewer floods are obviously expected (Figure 36A).

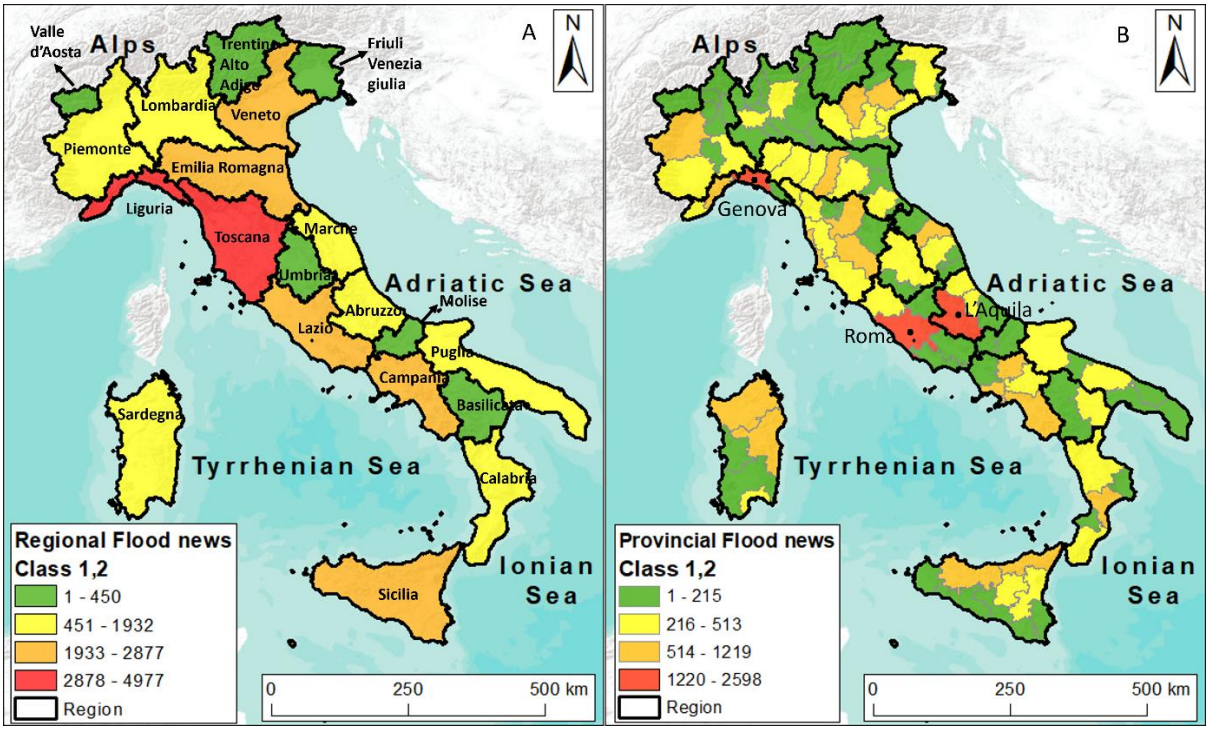

Figure 36: Spatial distribution of "Flood news": **A**) Regional and in **B**) Province aggregation with overall news (classes 1, 2). Genova is the province most affected by floods, followed by Roma and L'Aquila. Valle d'Aosta, Trentino Alto Adige and Friuli Venezia Giulia and the provinces along the South-East coast along Ionian Sea coast show a lower number of flood events. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

Figure 37 regards the relation between the news published near real-time (in class 1) or “Flood event” and the media impact or “Newspaper articles”, for each region.

Liguria is the first Region with important publication of articles for each flood event. Sicilia was the region with the highest number of “Flood events”. Lazio is the second region, regarding the number of “Flood events” and number of articles (“Newspaper articles”). While Toscana is the third in terms of “Flood events”, even if with important number of published articles. Molise, Basilicata and Valle d’Aosta are the last Regions with less data regarded flood events in agreement with the released articles.

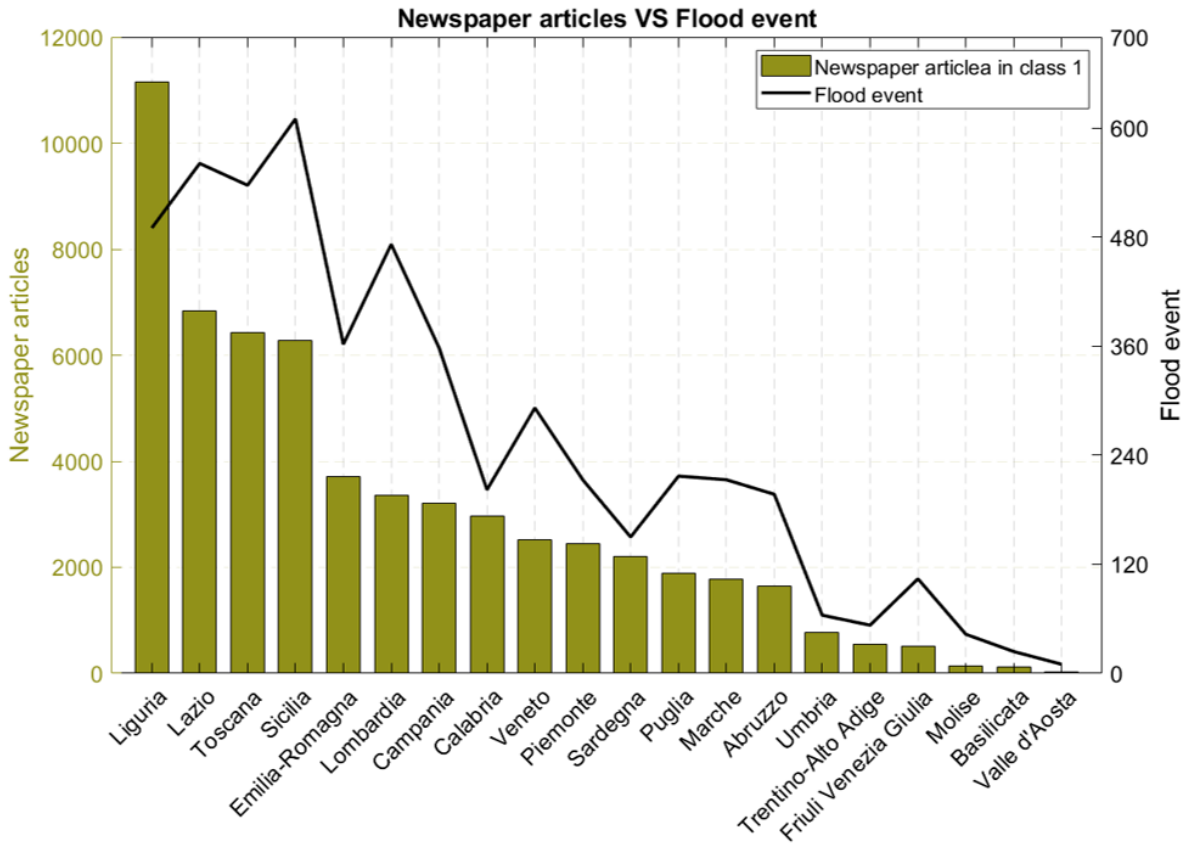

Figure 37: Regional distribution with comparison between the number of published articles or media impact respect to “Flood event”. The panel was generated using MATLAB R2021b.

The number of days with at least 1 reported flood event (“Flood day”) is higher in the western regions rather than in the eastern ones and in Sicilia the southernmost Region (Figure 38A). Overall, 5 regions out of 20 had at least 205 days with flood events, in the analysed period. Toscana, Lombardia, Lazio, Sicilia and Campania are the regions with the highest number of days characterized by floods. In particular, 247 days have been identified in Toscana, 238 in Lombardia, 234 in Lazio, 224 in Sicilia and 211 in Campania (Figure 38A). The Valle d’Aosta Region has shown the lowest number of days with flood: in this Region 10 events, distributed over 8 days, were present.

In a more detailed scale (Figure 38B), 7 provinces out of 107 have significant number of days with "Flood events" (56-193). For example, the Roma Province is characterized by 468 flood events, reported in 6398 articles, distributed over 193 days. The provinces that have less days with at least one flood event are located along north Italy in Valle d'Aosta Region and also in the northern in Piemonte.

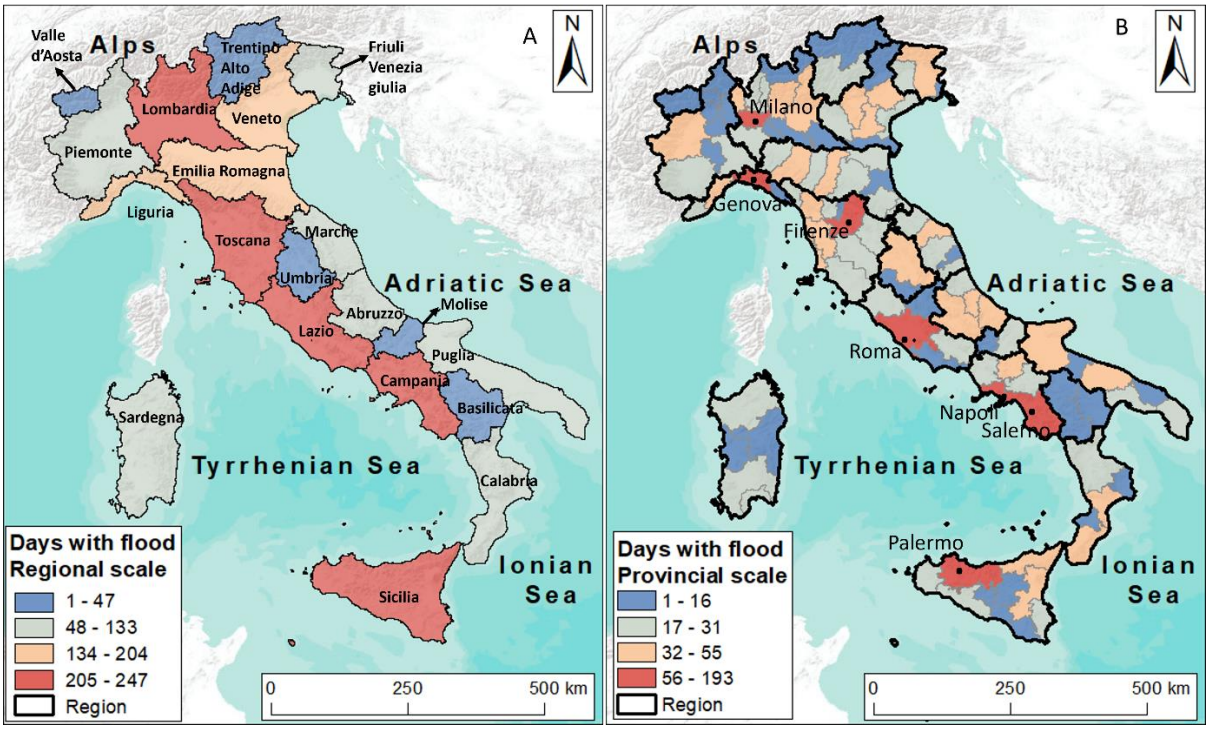

Figure 38: Spatial distribution of days with reported floods. **A** Regional distribution, **B** Provincial distribution. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

Temporal distribution

From a temporal point of view, the number of events showed a very sharp increase from 2016 (544 events) to 2018 (1125 events) (Figure 39A). The year with the highest number of floods-related articles was the 2014 with 12.588 media impact (Figure 39B).

Figure 39C displays a monthly distribution of the flood events identified by the class 1 data. November, October and September were the months more involved by floods. Indeed, November, in 10 years, reported 763 "Flood events" with 14.382 published articles (multiple articles can refer to the same "Flood event", as described in the previous section).

April, December and January were months with less events. For instance, in April 175 "Flood events" were reported by newspapers.

The number of days with at least 1 flood reported have been analysed (Figure 39D). The annual distribution follows a gradual increase of days with at least one flood from 2011 to 2014 with 1823

"Flood events" distributed over 516 days. From 2015 to 2018, 2932 "Flood events" have been collected, distributed over 727 days. The average increased passing to almost 3 days with flood to 4.

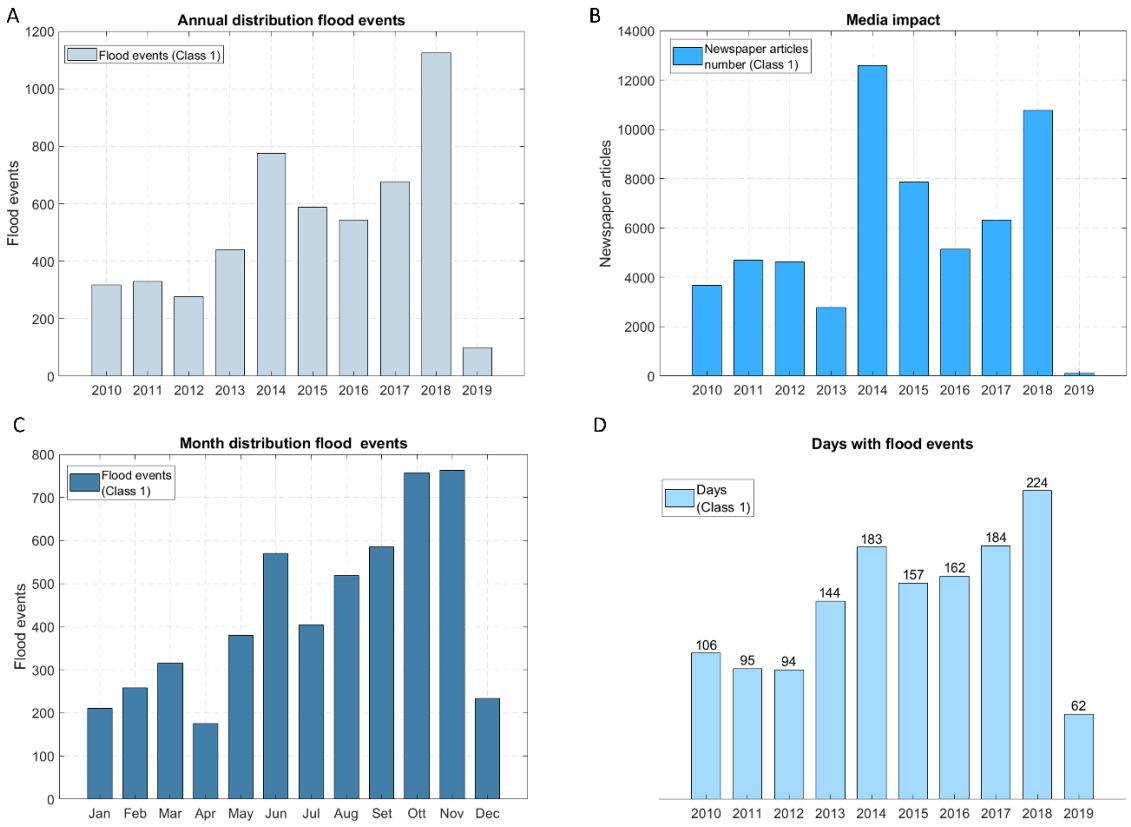

Figure 39: Temporal distribution of class 1 news. **A** "Flood events" annual distribution; **B** "Newspaper article" annual distribution; **C** monthly distribution of "Flood events"; **D** The number of days with at least 1 flood reported from 2010 to 2019. Panels were generated using MATLAB R2021b.

4.1.2.1 Correlation with traditional sensors

"Flood news", as well as "Landslide news" have been correlated with four datasets to get information about spatial and temporal distribution. Overall, the analysed data have been thus divided: 34.561 "Flood news" with 246.338 "Newspaper articles", 2040 rain gauges with 35.299 rainfall events, 99 data from Polaris and 1431 data from ReNDiS. All datasets cover the period 2010-2019, except the Polaris dataset, since it starts from 2011.

Spatial distribution

The "Flood news", as well as "Landslide news", cannot identify the exact location of an event, the maximum spatial results that can be obtained is the municipality. For this reason, the data have been grouped on regional base. The number of "Flood news" was used as a proxy to identify those areas

more affected by floods in the observed period, hence the most hazardous areas, while the number of ”Newspaper articles” was used as an estimator of intensity.

The most involved regions by flood news were mainly in the Central portion of the Country. Liguria and Toscana have been the Regions with the highest amount of “Flood news” (Figure 40A). Liguria and Toscana have been also the Regions with the highest correlation between variables, considering relevant rainfalls distribution (Figure 40C), IDMS (Figure 40D), affecting region (Figure 40E) and earmarked funds (Figure 40F), but not rainfall data (Figure 40B). Friuli Venezia Giulia, Lombardia and Piemonte resulted to be the rainiest Region of Italy. IDMs distribution resulted more fragmented across the Country. For example, Calabria, Toscana, Veneto, Sardegna and Sicilia were the Regions with significant IDMs number, respectively 27, 27, 26, 24 and 24. Only Molise and Valle d’Aosta showed 0 IDM after a flood.

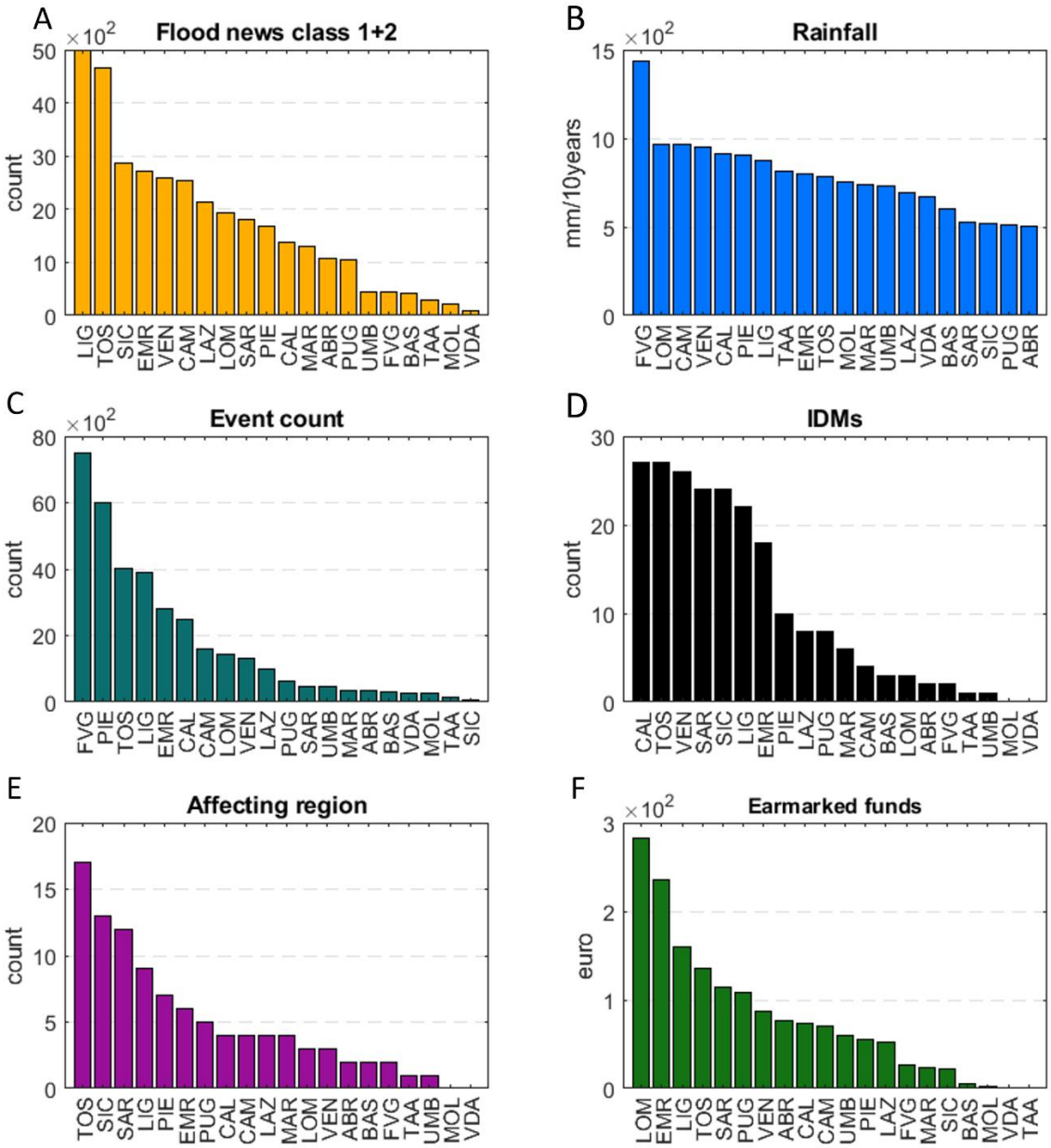

Figure 40: Spatial distribution with different information: **A** "Flood news" and "Newspaper articles" from Social Media; **B** Rainfall data with mm/y and in **C** events count; **D** Polaris with IDEMs (injured, death, evacuated and missing) number; **E** involved Region; **F** earmarked funds for the soil protection (euro – ReNDiS with focus for better vision). **ABR**: Abruzzo, **BAS**: Basilicata, **CAL**: Calabria, **CAM**: Campania, **EMR**: Emilia-Romagna, **FVG**: Friuli-Venezia Giulia, **LAZ**: Lazio, **LIG**: Liguria, **LOM**: Lombardia, **MAR**: Marche, **MOL**: Molise, **PIE**: Piemonte, **PUG**: Puglia, **SAR**: Sardegna, **SIC**: Sicilia, **TOS**: Toscana, **TAA**: Trentino-Alto Adige, **UMB**: Umbria, **VDA**: Valle d'Aosta, **VEN**: Veneto. The arrow indicates the increasing direction of allocated funds. Panels were generated using MATLAB R2021b.

In general, the distribution of "Flood news" (Figure 41A) and "Newspaper articles" (Figure 41B) were resulted not coherence with rainfall data (Figure 41C). Clearer is the case of Friuli Venezia Giulia. Friuli Venezia Giulia presented low values of "Flood news" and "Newspaper articles", but important values of rainfall data. The news variable is in agreement with the count event or relevant rainfalls (Figure 41D), IDMs (Figure 41E) and earmarked funds (Figure 41F).

Valle d'Aosta, Molise, Umbria and Marche have been the Regions that shown the best coherence between variables. In this sense, as for the landslides, the morphology of the territory, the climatic conditions, the density of people and buildings at risk can bias the distribution of flood events.

In conclusion, Valle d'Aosta, Piemonte, Liguria, Toscana, Emilia Romagna, Umbria, Marche, Abruzzo, Lazio, Molise, Basilicata and Trentino Alto Adige have been the Regions that shown higher coherence between the datasets. Vice versa, Lombardia, Veneto, Campania, Sicilia, Puglia, Calabria and Sardegna have been the Regions with lower coherence.

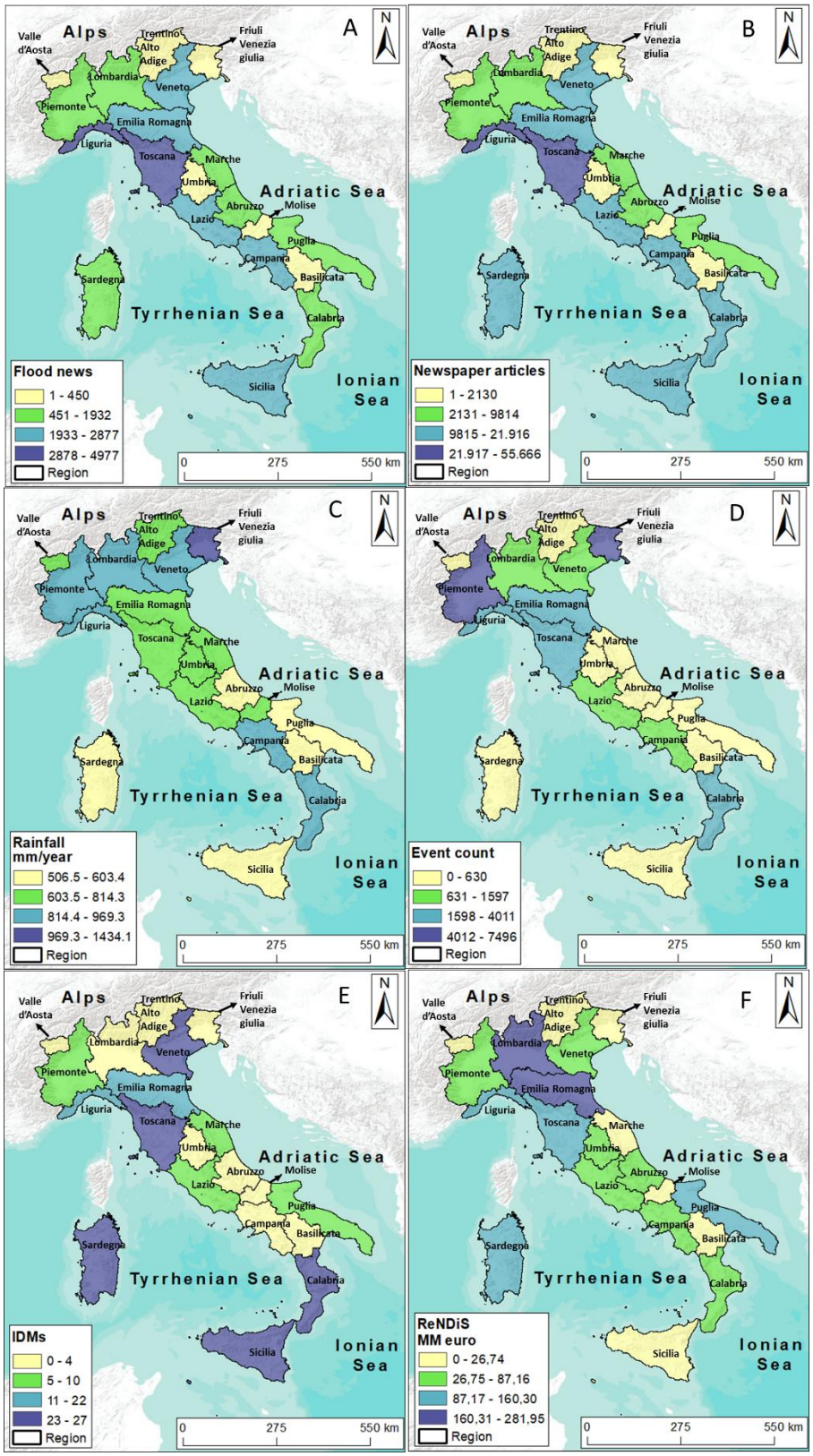

Figure 41: **A** Regional aggregation with "Flood news", estimating flood hazard; **B** Regional aggregation with mediatic impact, estimating the flood intensity; **C** Regional distribution with rainfall data for 10 years (mm/10 years); **D** Regional aggregation of relevant rainfall events count. Events display relevant differences about spatial distribution; **E** Regional aggregation with IDMs from Polaris dataset. The period covers only 9 years, from 2011 to 2019; **F** Regional aggregation with earmarked funds for the soil protection from ReNDiS (euro/10 years) for 10 years, considering flood events. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

Temporal distribution

By a temporal a point of view, each Italian Region experienced some floods in the investigated period, with approximately 3821 IDEMs per year.

As well as “Landslide news”, “Flood news” showed an increasing trend from 2010 to 2014, which repeats in the period 2015-2019. At the end of the first quinquennium, the 2014 featured by the highest number of “Flood news” (3146) with a mean rainfall of 1007,6 mm/year (Figure 42A-B). In this year, 11 regions out of 20 were involved by floods with 4706 IDEMs (Figure 42E-F) as consequences. In contrast, the 2014 was the third year with less earmarked funds for soil protection, almost 9 million of euro, behind only to 2016 with 8 million of euro (Figure 42F) and to 2012 with absence of funds.

In the second quinquennium, the year with the highest number of “Flood news”, with 7159 data, was the 2019. Although the rainfall data were constants respect to the previous year, the 2019 was also the second year with significant events count with 4529 relevant rain events (Figure 42A-B-C). In the same year, 11 regions were affected by floods with 4529 IDEMs (Figure 42D-E). The 2019, in these five years, it was the second year with important earmarked funds, almost 114 million euros (Figure 42F) behind only to 2015 with 319 million of euros.

In general, the “Flood news” showed an increase from 13.343 news in the period 2010-2014 to 21.217 in 2015-2019 as well as for IDEMs from 17.355 to 20.855 (Figure 42D). Vice versa for the rainfall events (Figure 42B), relevant rainfall events or events count (Figure 42C) and “Newspaper articles” the trend resulted inverse. For example, “Newspaper articles” were featured by 134.318 articles for the first quinquennium to 112.020 in the second (Figure 42A).

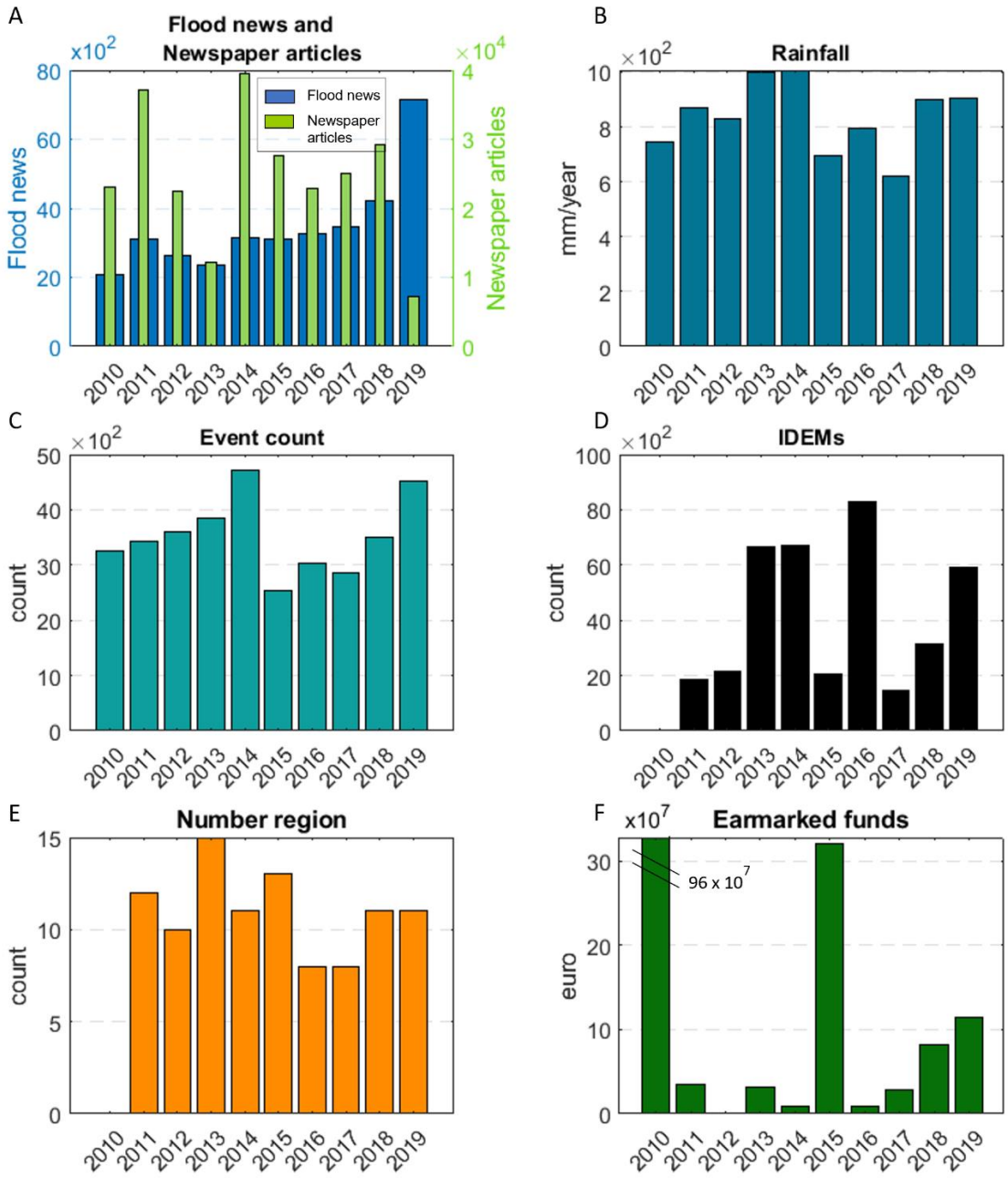

Figure 42: Temporal distribution with different information: **A** "Flood news" and "Newspaper articles" from Social Media; **B** Rainfall data with mm/y and in **C** events count; **D** Polaris with IDEMs (injured, death, evacuated and missing) number; **E** involved Region; **F** earmarked funds for the soil protection (euro – ReNDiS with focus for better vision). Panels were generated using MATLAB R2021b.

4.1.2.2 Text analysis and word distribution

From 2010 to 2019, 34.560 news have been gathered by the used data mining algorithm.

The dataset provides the headline as the only textual source. In the beginning, the text has been pre-processed by removing all those textual parameters lacking literary meaning within the sentences: articles, punctuation, special characters, number of words with below 2 letters, low word frequency. Once the text was cleaned, two analyses have been applied: qualitative analysis with word cloud (Figure 43A) and quantitative analysis with word frequency count (Figure 43B).

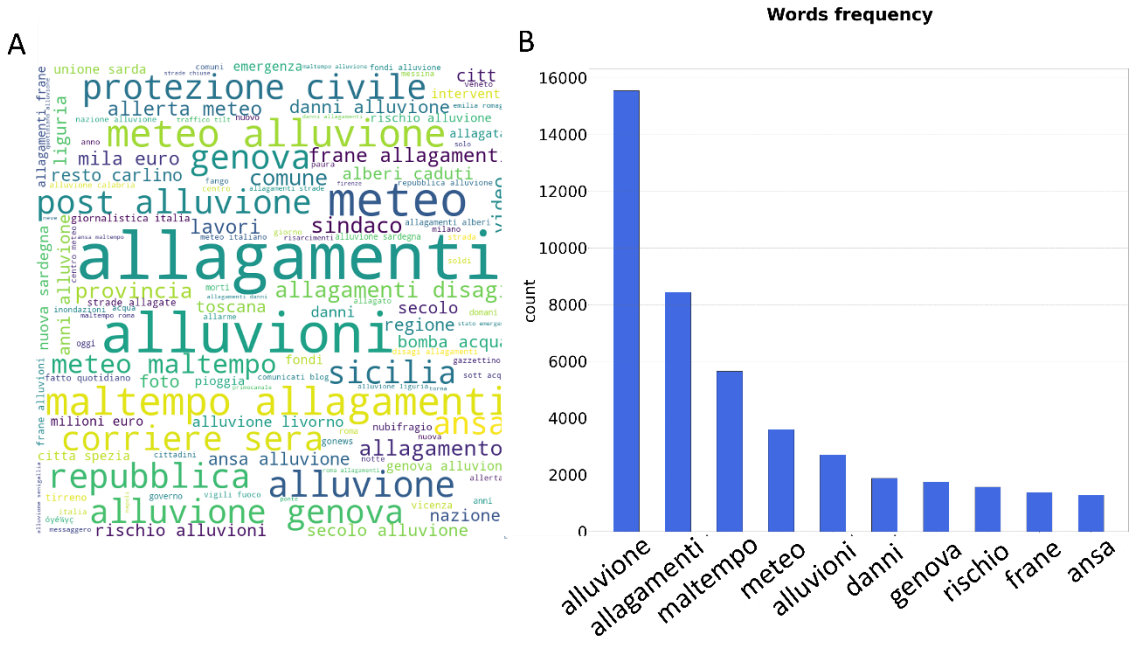

Figure 43: Qualitative representation of LDA using Wordcloud in **A** and in **B** a quantitative representation with the words frequency with the first ten words. Panels were generated using Python. Below the translation for each word; alluvione: flood, allagamenti: flooding, maltempo: bad weather, meteo: weather, alluvioni: floods, danni: damage, Genova: Genova, rischio: risk, frane: landslides, ansa: newspaper name ANSA.

Among them, 5172 news had useful information about the geo-localization and the date of flood event. According to the adopted classification criteria, the identified news has been classified as follows:

- Class 1: 5172 news (15%) (Figure 44A).
- Class 2: 29.388 news (85%) (Figure 44A).

The Class 3 wasn't identified. Textual analysis has been conducted to retrieve the frequency of words inside the headlines. In Figure 44B-C the most frequent words of the headlines of the class 1 and 2 news are reported, respectively. The term "flood" is present in all categories, linked often to "bad weather".

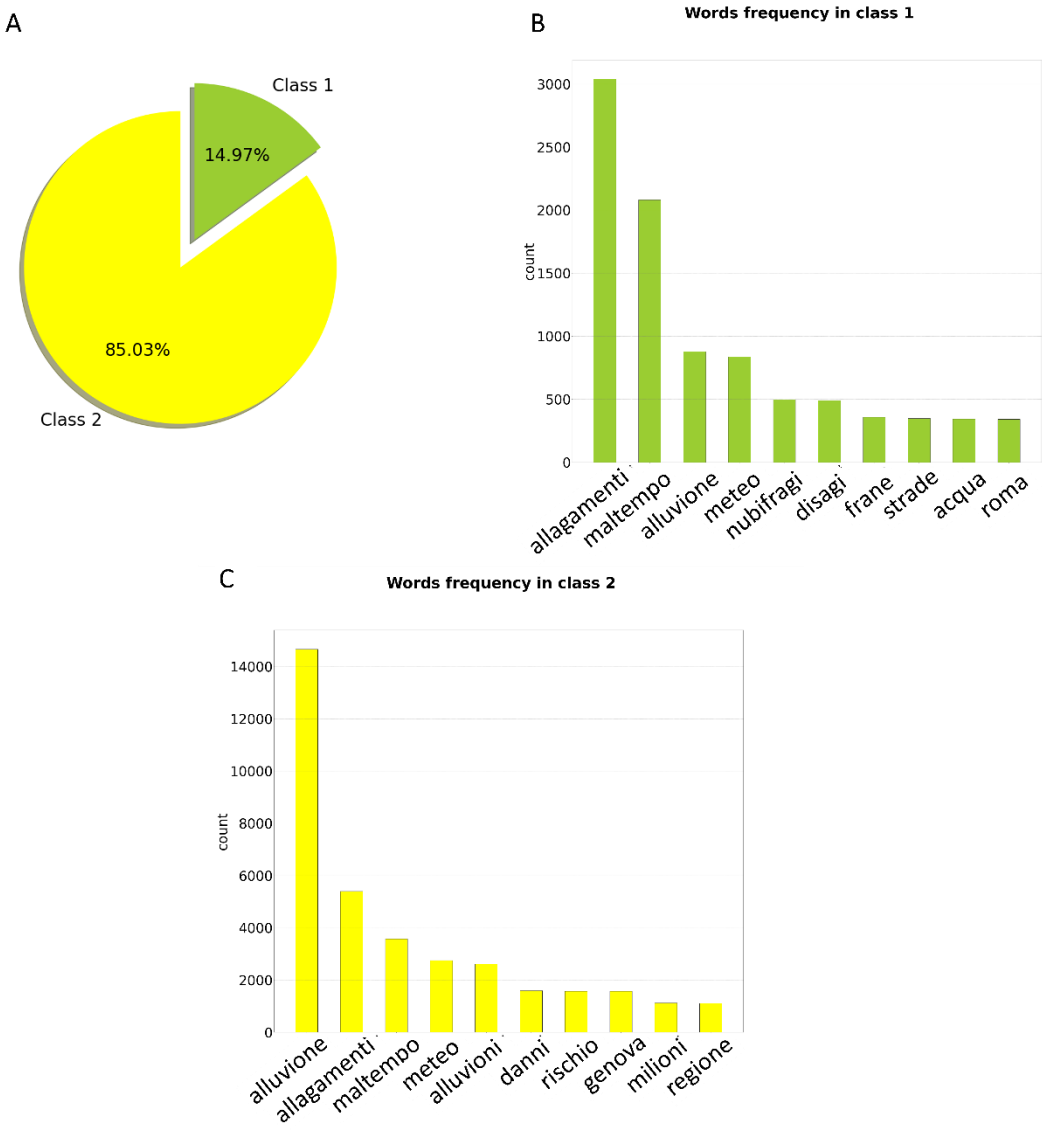

Figure 44: **A** Overall landside news classification. **B** Words' frequency in the headlines inside Class 1, **C** Words' frequency in the headlines as inside Class 2. Panels were generated using Python. Below the translation for each word; allagamenti: flooding, maltempo: bad weather, alluvione: flood, meteo: weather, nubifragi: storms, disagi: inconvenience, frane: landslides, strade: roads, acqua: water, roma: Roma, alluvioni: floods, danni: damage, rischio: risk, Genova: Genova, milioni: milions, regione: region.

4.2 Data mining for tweets dataset

Tweets from traditional and internet media offer a variety of information types about affected individuals and messages of caution and advice. Media are also the most prominent source of information regarding infrastructure and utilities (Olteanu et al.,2015). In this project, 9 slots have been

harvested using the data mining technique. From Twitter, 13.350 data points were harvested from 2011 to 2019 (Table 15 and Figure 46).

year	from	to	data
2011	22/03/2011	22/03/2011	1
	01/10/2011	30/11/2011	420
2012	01/09/2012	31/12/2012	693
2013	01/01/2013	31/05/2013	1028
2014	01/01/2014	31/05/2014	1747
	01/07/2014	30/11/2014	1319
2015	22/02/2015	26/02/2015	1626
2016	24/11/2016	28/11/2016	1656
2017	05/08/2017	08/08/2017	486
2018	28/10/2018	31/10/2018	2273
2019	24/11/2019	24/11/2019	2100

Table 15: Nine slots extracted within Twitter.

According to the adopted classification criteria, the identified news has been classified as follows:

- Class 1: 4805 (36%) (Figure 45A and B). Not all landslide-related tweets contained location information; there, tweets not referring to a location were not used in the subsequent spatial analysis. In total 526 data show no coordinate within the text.
- Class 0: 8544 (64%) (Figure 45A and B) not providing any information.

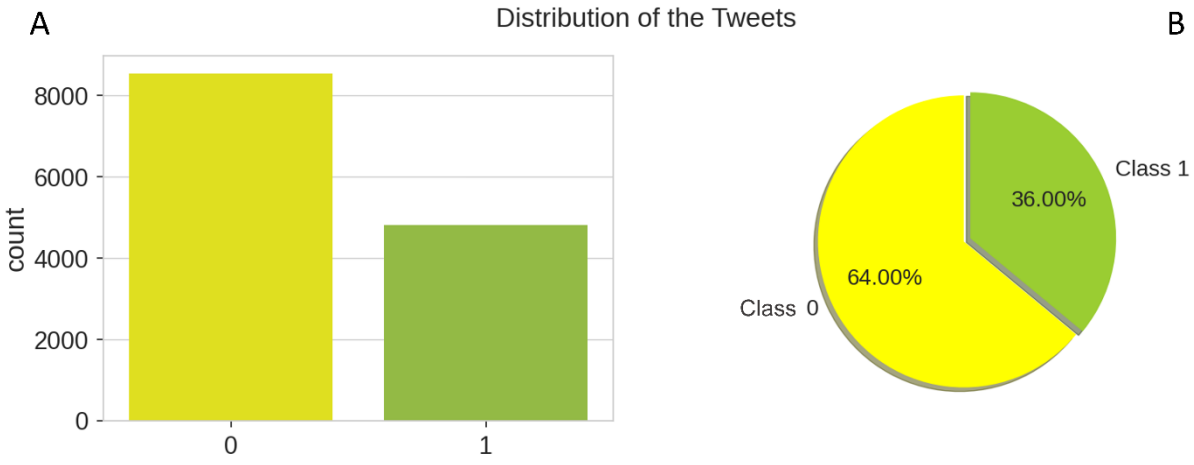

Figure 45: Considering only the label describing the landslide event, in **A,** the distribution of data for each target is shown, and in **B,** the distribution is expressed as a percentage Panels were generated using Python.

In class 1, 4158 tweets were assigned approximate coordinates based on specifics within the tweet text, such as municipality, region, and street. Figure 46 displays the spatial distribution of each obtained dataset from Twitter.

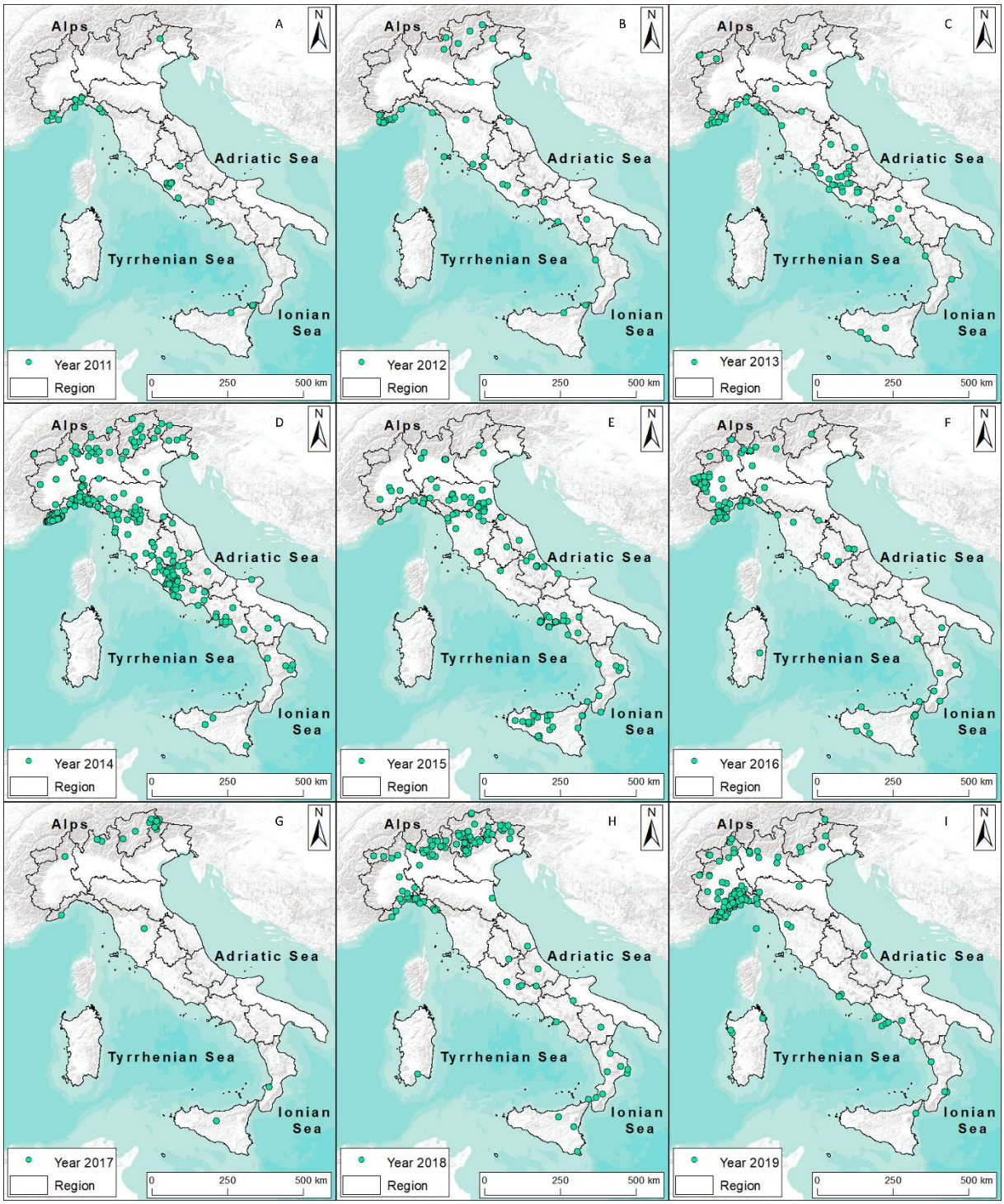

Figure 46: Nine maps using coordinates from text are shown as follows: in **A,** data from some days or months during 2011; **B,** data from some months during 2012; **C,** data from some months during 2013; **D,** data from some months during 2014; **E,** data from some days during 2015; **F,** data from some days during 2016; **G,** data from some days during 2017; **H,** data from some days during 2018; **I,** data from some days during 2019. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

Figure 47A presents the total dataset of Twitter using coordinates from the tweet text. The distribution of tweets did not continue for 10 years. However, as demonstrating by Wang et al., (2021) Twitter can be used to outline which region had the most resilience in terms of tweet publications.

Analysing the tweet counting, the Liguria Region presents the highest values, followed by Lazio and Calabria. The regions with the lowest counts are Marche and Puglia. Molise does not display user interactions (Figure 47B). Considering retweets or the media impact of events, the first five regions with significant interactions between users are Liguria, Piemonte, Calabria, Veneto and Trentino Alto Adige. Molise and Puglia show the lowest values of retweets (Figure 47C).

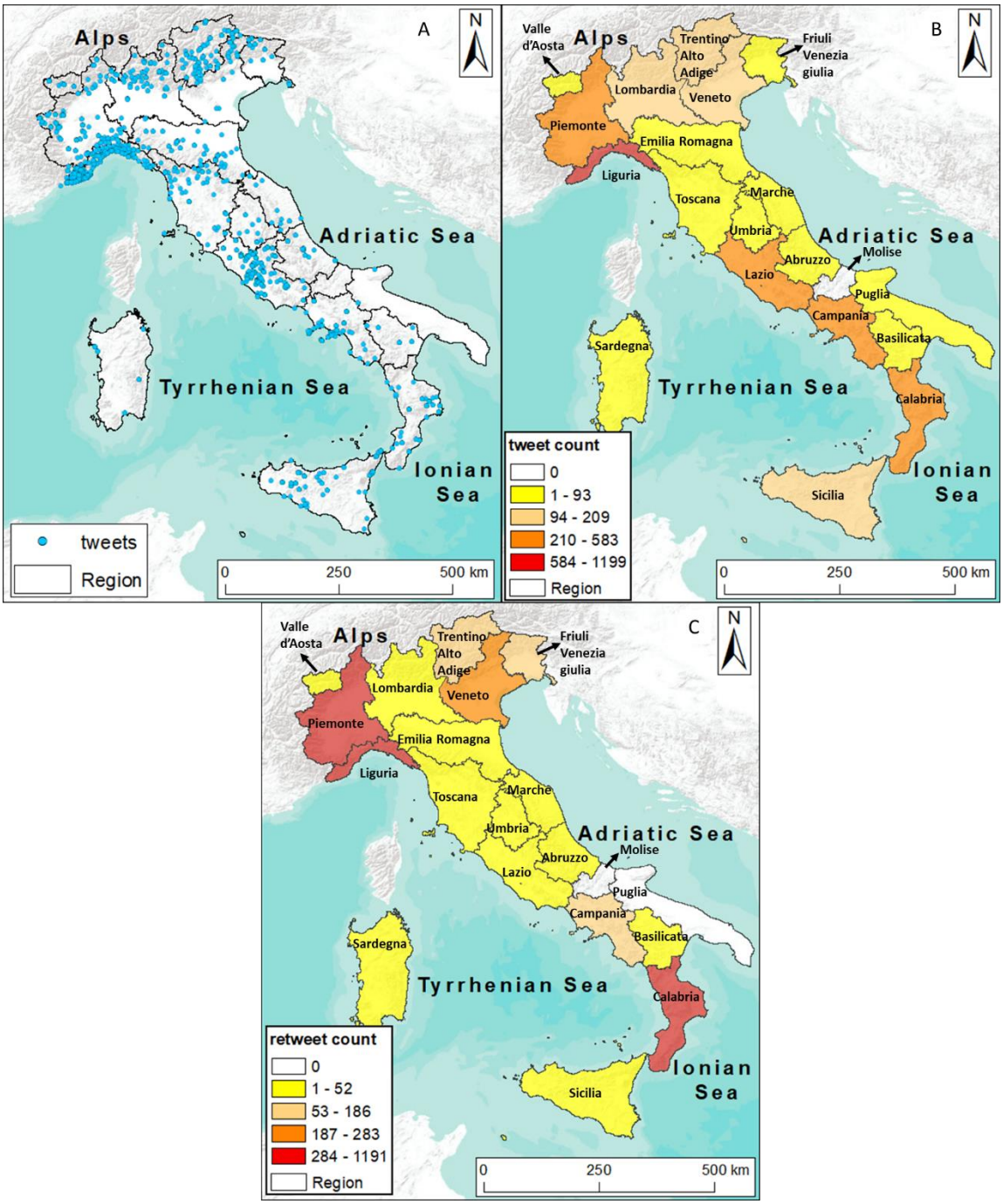

Figure 47: Considering only the label describing the landslide event, in **A** the distribution of data in class 1 is shown, in **B** the distribution of tweet count for each region, and in **C** the retweet distribution for each region. Such a parameter can present the media impact of the event and interaction between users during some events within Italian territory. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

Data with further information, such as street number, village or picture information (photo/video), allow a significant localization and identification of the landslide event. Summarising this information of photointerpretation, it was possible to assign accurate coordinates to the data and, hence, to the landslide event.

Within class 1, 1529 tweets were allocated the right coordinates. (Figure 48A). Considering the tweet counting, Liguria, Campania and Emilia Romagna are the regions with significant publication values. Molise, Puglia and Umbria did not show tweets (Figure 48B). Regarding the interaction between users or retweets, Liguria, Campania and Piemonte were the first three regions with important values. Umbria, Abruzzo, Molise and Puglia did not present interactions between users.

Figure 48: Considering only the label describing the landslide event, in **A,** the distribution of data in class 1 using the true coordinates of the event is shown. The event can be identified through photo, video, address within tweet text, photo interpretation, etc. In **B,** the distribution of count tweets and Liguria show the highest values of tweet publishing. In **C,** the distribution of retweets is presented, where Liguria and Campania display significant values of interaction between users during an event and, consequently, make an important media impact. The maps were generated using ESRI ArcMap 10.8.1 (https://www.arcgis.com/home/item.html?id=33064a20de0c48d2bb61efa8faca93a8).

4.2.1 Exploring dataset

Before applying the deep learning technique, some analyses were applied to obtain information about the dataset from Twitter. Natural language processing techniques were adopted to outline the word distribution for each target (0 and 1) and, further, at the same time alternating the application of the preprocessing or cleaning data.

News and tweets have been compared to outline similar trends during each period. Targets with landslide information (for news class 1+2 and tweets in class 1) and without information have been correlated. A satisfactory relationship was verified with the first case (data with landslide information), while a poor relationship was verified with the second data (without information), particularly with tweet data.

Finally, three case studies were investigated, in particular Emilia Romagna, Campania and Liguria. One focus has been applied to the last case study. The Liguria Region involved intense rainfall in November 2019, which triggered several landslides in the territory. One landslide caused a fallen viaduct. News and rainfall data were correlated to a tweet dataset for this specific event.

4.2.2 Some analysis of natural language processing

In the beginning, the tweet text was preprocessed by removing all textual parameters lacking literary meaning within the sentences: HTML, stop words, converted @username to AT_USER, tickers, number, lowercase, hyperlinks, hashtags, punctuation, number of words with fewer than 3 letters, whitespace (including new line characters), and space remaining at the front of the tweet.

Once the text was cleaned, qualitative analysis was applied to obtain the word cloud (Figure 49A) and quantitative analysis for word frequency counting (Figure 49B).

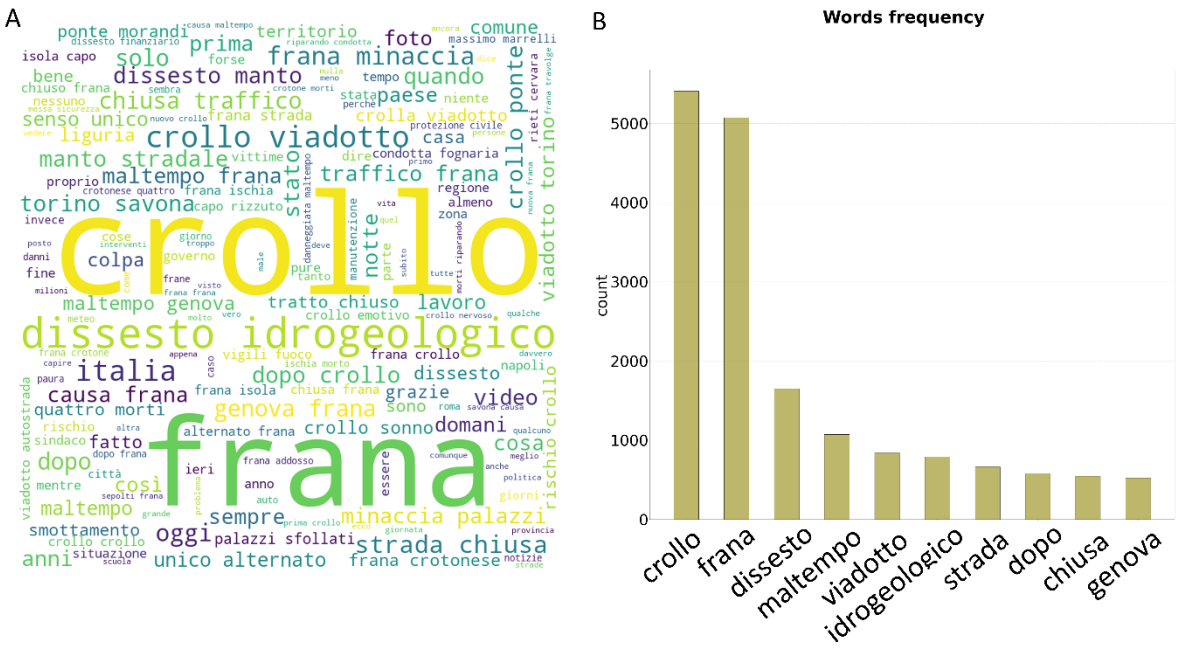

Figure 49*:* In **A** a qualitative representation of the frequency of words within the tweet text; in **B** the frequency of words. Panels were generated using Python. Below the translation for each word; crollo: collapse, frana: landslide, dissesto: instability, maltempo: bad weather, viadotto: viaduct, idrogeologico: hydrogeological, strada: road, dopo: next, chiusa: locked.

There are other basic strategies for text analytics and more advanced techniques that leverage machine learning, statistical and linguistic techniques. Some of the techniques can be used to begin investing in text analytics. Common text analytics techniques include word frequency. This is a technique used to measure the most frequently occurring words and phrases in specific conversations. For instance, you could use text analytics and word frequency to determine which features citizens mention most often during an event. Collocation and concordance can help to identify the words that usually occur at the same time and context of those. The common type of collocation is bigrams (Figure 50). Bigrams made up two adjacent coexisting words: ‘time table’, ‘air conditioner’ or ‘ice cream’. This technique helps to identify semantic structures (semantic means words connected with a meaning), and it counts bigrams and trigrams as one word.

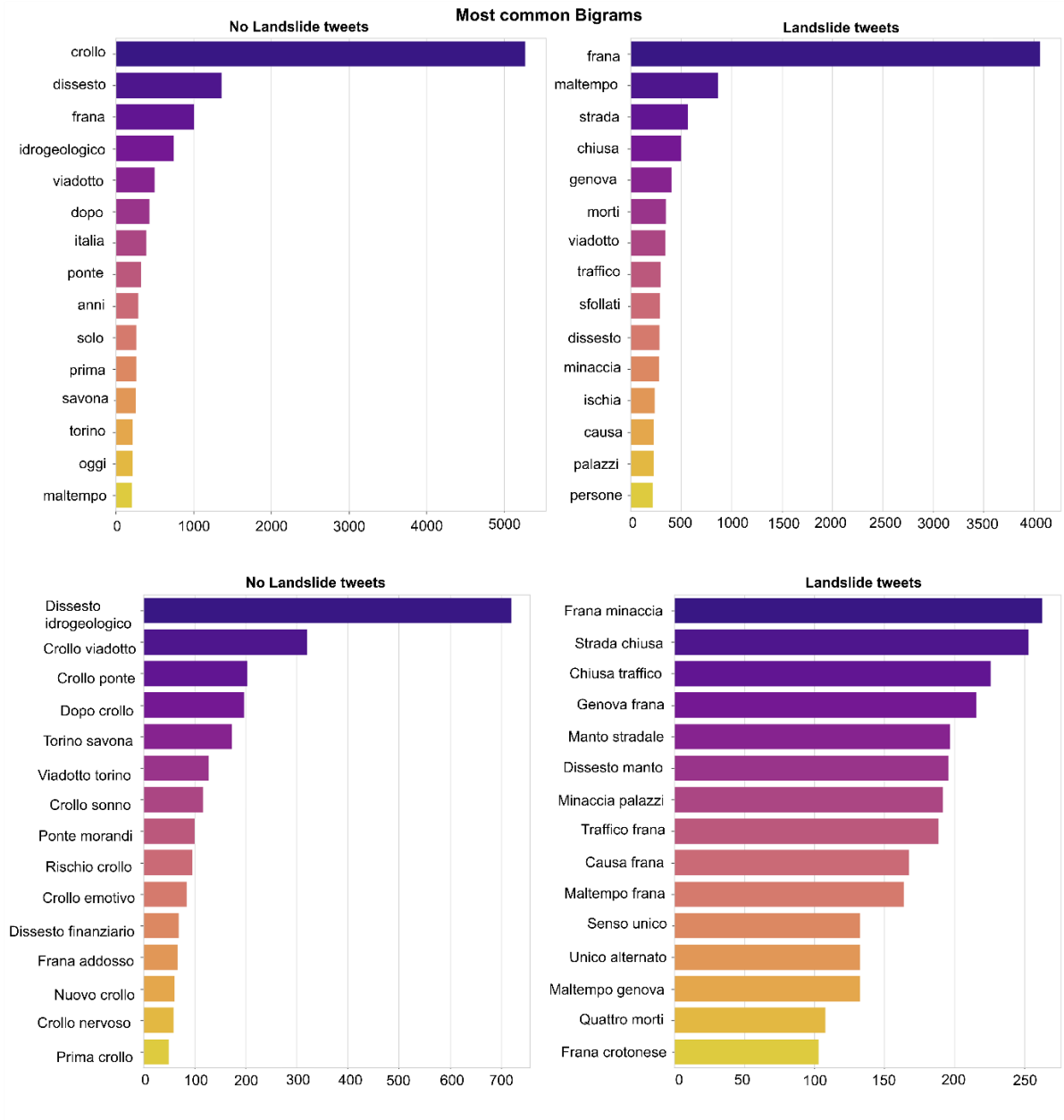

Figure 50: Two bigrams were used to obtain information about word distribution for each tweet respect target. The first two bigrams consider only one word distribution, and the second bigram describes the frequency of co-occurrence between the two words. This technique was applied after preprocessing. Panels were generated using Python. See the Appendix after the bibliography for the translation.

Creating a "bag-of-words" model it is possible to create a co-occurrence network. A co-occurrence network is an undirected graph, with nodes corresponding to unique words in a vocabulary and edges corresponding to the frequency of words co-occurring in a document. Use co-occurrence networks to visualize and extract information on the relationship between words in a corpus of documents. In summary, it is possible to discover which words commonly appear with a specified word. In this case, four networks are displayed in Figure 51. Three datasets have been considered: i) generic classified dataset; ii) data classified in class 1 (or "Landslide tweets") and iii) data classified in class 0 (or "No

Landslide tweets). To create these networks have been sampled randomly 100 data for each dataset, for better to represent the distribution of frequencies. Figure 51A and B display the relationship between the words in general for 100 random data in whole dataset. Figure 51A point out the strong correlation between “frana”, “crollo”, ”strada”, “maltempo”, “sicurezza”, “morti” and “sfollati” (see the Appendix after the bibliography for the translation.). Figure 51B highlights the trend considering “frana” the stop point of frequency. Figure 51C and D show the distribution of words around the main term “crollo” for class 0 and “frana” for class 1. The words have been chosen on the base of preview frequency analysis in Figure 50. The nearest words of “crollo” for the class 0 are “strada” to sign damages caused by different accidents (natural, structural etc..) and then there is an equal distribution with other terms not useful to identify the landslide event. In fact, often the word “crollo” is associated with political/economic situation (“governo” and “borsa”), sentimental (“emotive”, “notte” and “domani”) or generic without other specifics (“strada”, “viadotto, “rischio”, “totale”, ”visto” and “situazione”) (Figure 51C- see the Appendix after the bibliography for the translation.). While, the closest words of “frana” are “maltempo” and its synonyms #maltempo, “provinciale” (referring to the street), the synonyms “#frana”, but also “stradale” and “palazzi” to highlight the involvement of infrastructure (Figure 51D- see the Appendix after the bibliography for the translation.).

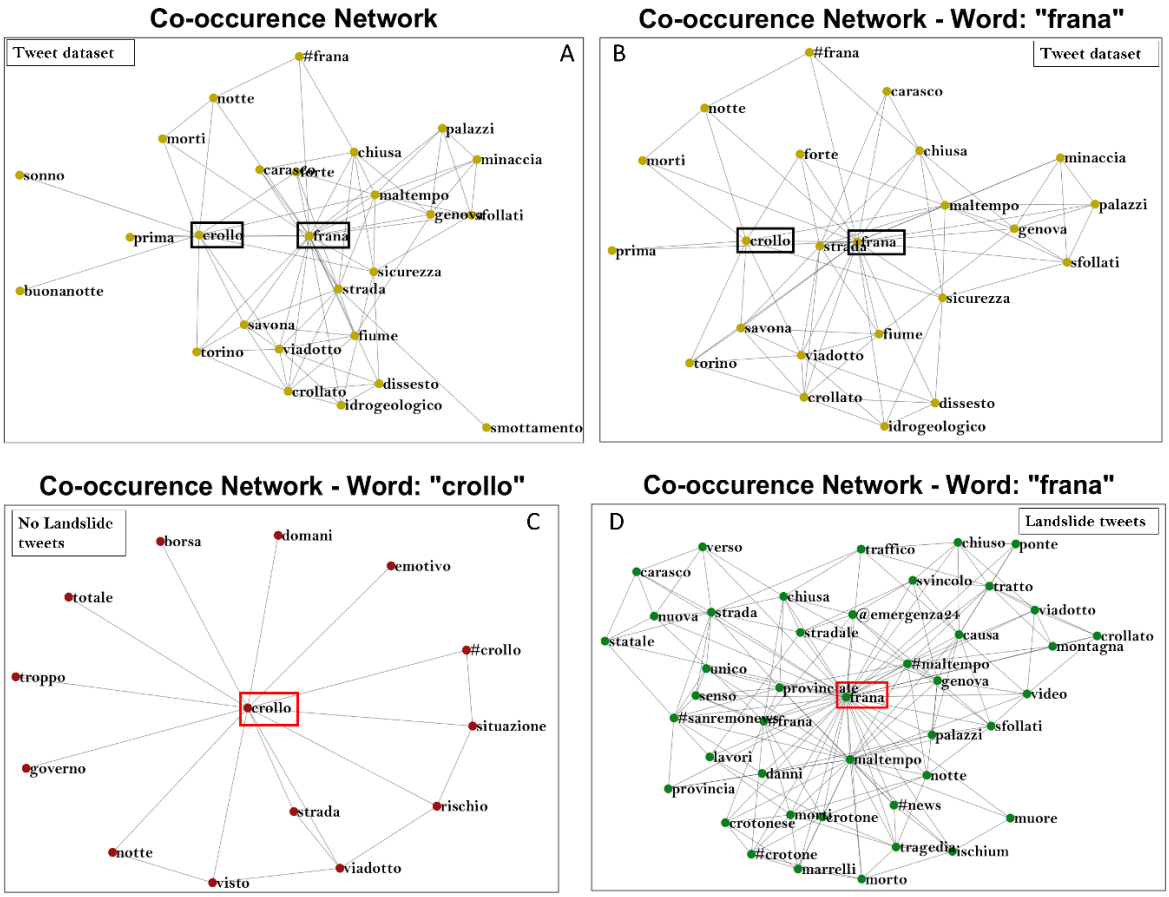

Figure 51: Co-occurrence considering 100 random data for all datasets in **A**. In **B**, the word 'Landslide' was considered as a keyword to obtain the frequency of words related to it. In **C**, the keyword 'collapse' was chosen to obtain co-occurrence. Again, 100 data were considered randomly. 100 data were selected for each analysis and thus for each representation for reasons of visualisation and clarity of presentation. In **D**, 100 pieces of data were randomly considered for class 1 or 'Landslide event'. The word 'landslide' was chosen as the linking keyword based on the frequency of the words in the previous results. Panels were generated using MATLAB R2021b. See the Appendix after the bibliography for the translation of some words.

A similar procedure was applied to each target 0 (“No Landslide tweets”) and 1 (“Landslide tweets”) to determine the word count within tweets and their frequency (Figure 52). The first two panels above compare the word count and frequency of the tweets without applying any data cleaning. An even distribution of words was found in “Landslide Tweets” compared to “No Landslide tweets”. In the “No Landslide tweets”, a greater incoherence and distribution in count and frequency was shown.

By applying to preprocess, the last two panels of Figure 52 have been derived. The word count consequently decreases, but the frequency remains high in “Landslide tweets” compared to “No Landslide tweets”.

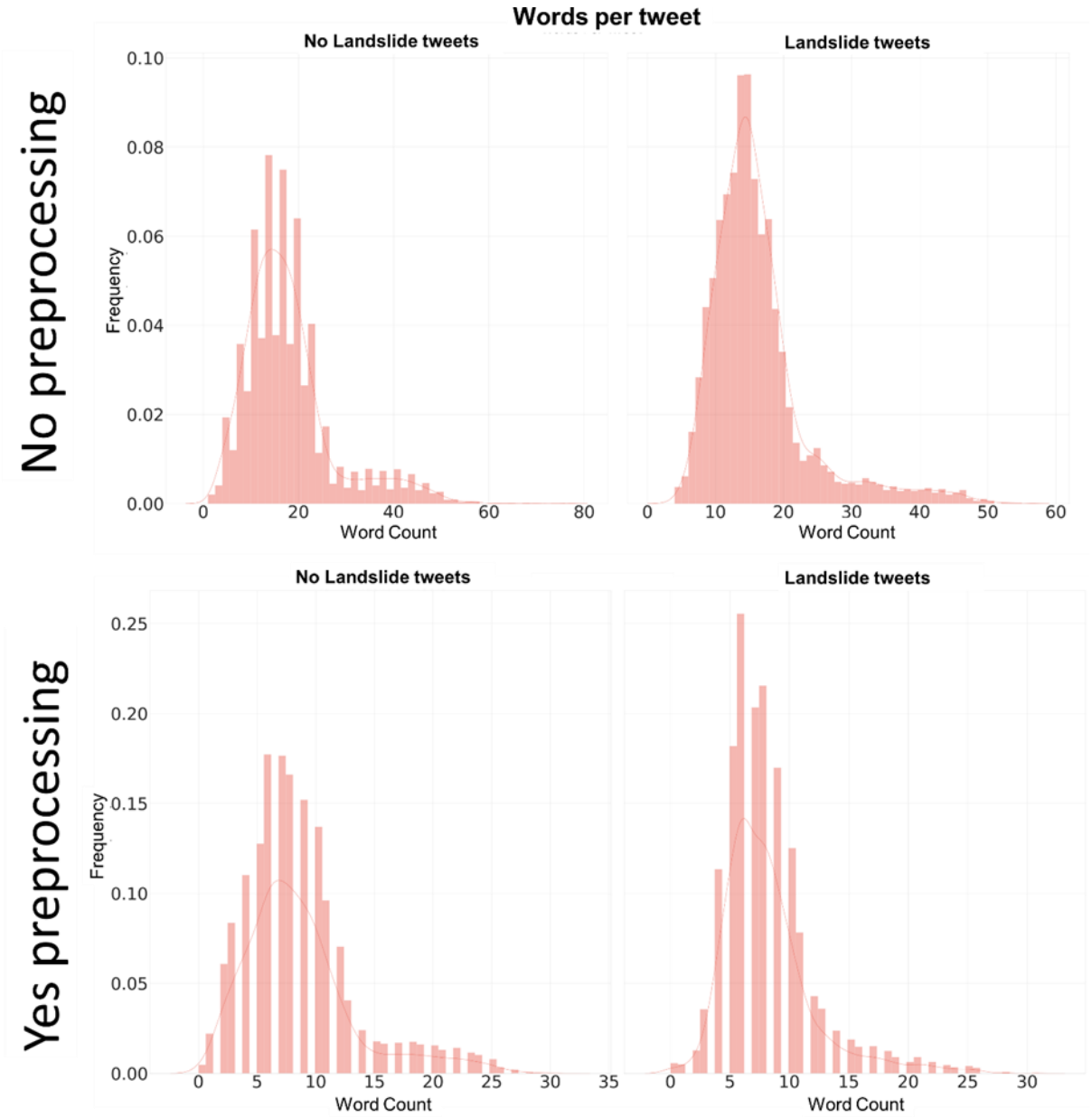

Figure 52: Four panels show the word count per tweet and their frequency per target of "No landslide" tweets and landslide tweets. No preprocessing was applied in the first two top panels. Preprocessing was applied in the bottom two panels. The distribution of words in the last two cases decreased, but the frequency increased accordingly. Both cases show a larger breakdown in the number of words for tweets not describing the landslide event than for Landslide tweets reporting the event. Panels were generated using Python.

4.2.2.1 Comparison between news and tweets

A comparative analysis was carried out between the news dataset and the many slots of tweets. Plot graphs were considered (Figure 53) due to the large extraction radius for 2011, 2012, 2013, and 2014. For datasets featuring few days or small slots, daily data were regarded for 2015, 2016, 2017, and 2018. A separate discussion was carried out for 2019, including single-day data that will be referred to in the next chapter.

The first analysis concerned the comparison between the news in class 1+2 and the tweets in class 1. Class 1 tweets do not consider the temporality of the event, i.e., whether the landslide occurred days ago or in real-time. For this reason, they were compared with the news classified into classes 1 and 2. Figure 53A, C, E, and G illustrate the predominance of newspaper news publications over tweets. On the other hand, there is a good correspondence of publication timing.

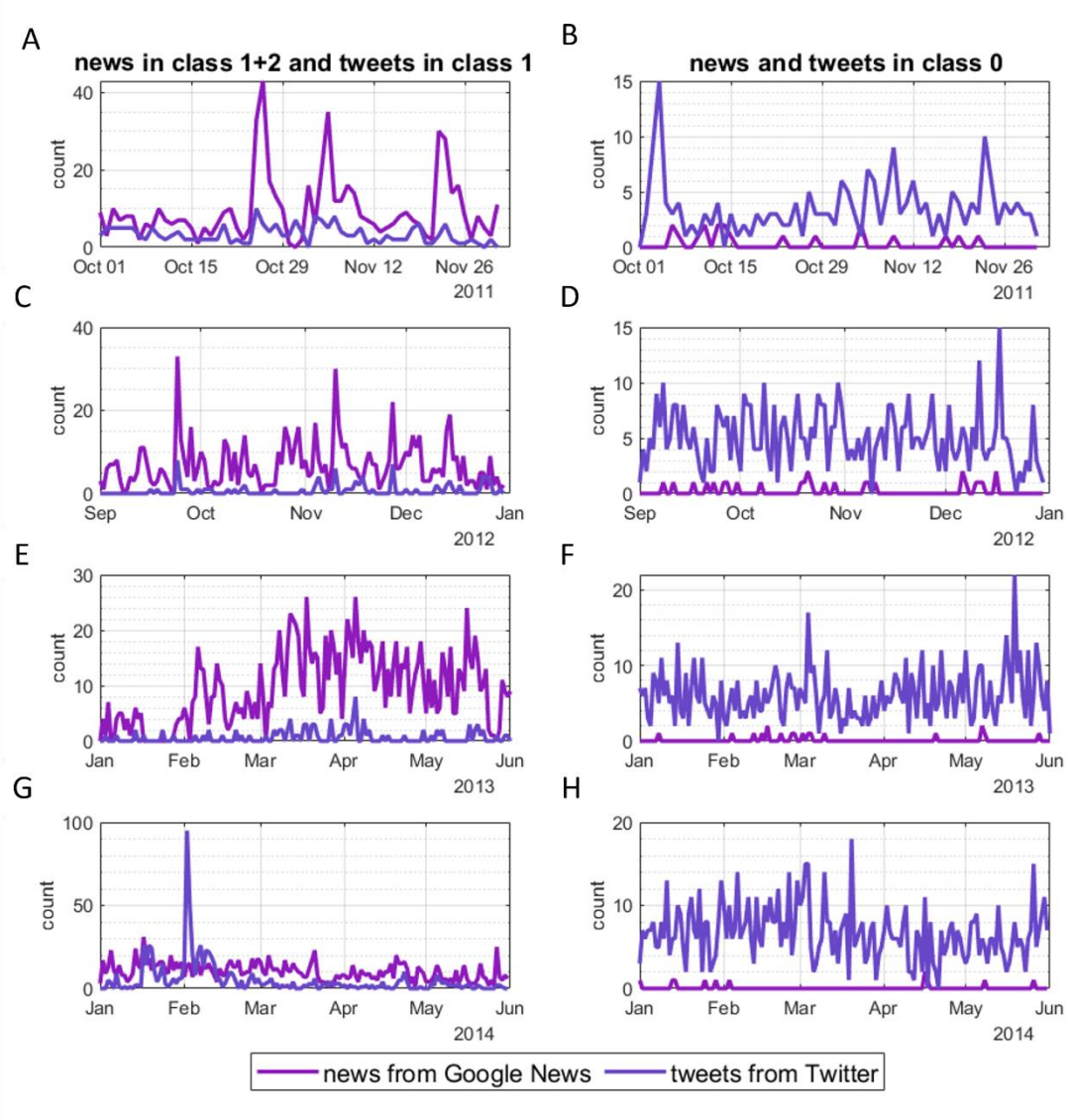

Figure 53: In the first column (**A**, **C**, **E** and **G**), news in class 1+2 compared with tweets in class 1 for each year from 2011 to 2014. The time distribution ranged from 5 months to 6 months. In the second column (**B**, **D**, **F** and **H**), news in class 3 and tweets in class 0, data that did not describe landslide events, have been compared. The time distribution covers the same year from 2011 to 2014. Panels were generated using MATLAB R2021b.

Poor correspondence was visible between news in class 3 and tweets in class 0 (Figure 54B, D, F and H). Undoubtedly, in previous results, it was already clear that there was a small number of wrong news items collected because the news was already filtered by the SECaGN system. Hence, many news items

not related to the event were rejected. Conversely, many tweets were classified as incorrect, which anticipates the strong heterogeneity of Twitter texts.

Extractions ranging from 4 to 5 days were carried out from 2015 to 2018. Figure 54A, C, E and G represent the inherent landslide data. All years have a very good correlation between news and tweets; see 25/02/2015, 27/11/2016, 05/08/2017 and 28/10/2018. On the day after the event, there is a clear decrease in the publication of tweets, while the publication of news remains sizeably high (e.g., 26/02/2015, 28/11/2016, 06/08/2017 29/10/2018). Figure 54B, D, F and H show the performance of the respective 0 classes. As before, there is a clear preponderance of erroneous tweets compared to Google News.

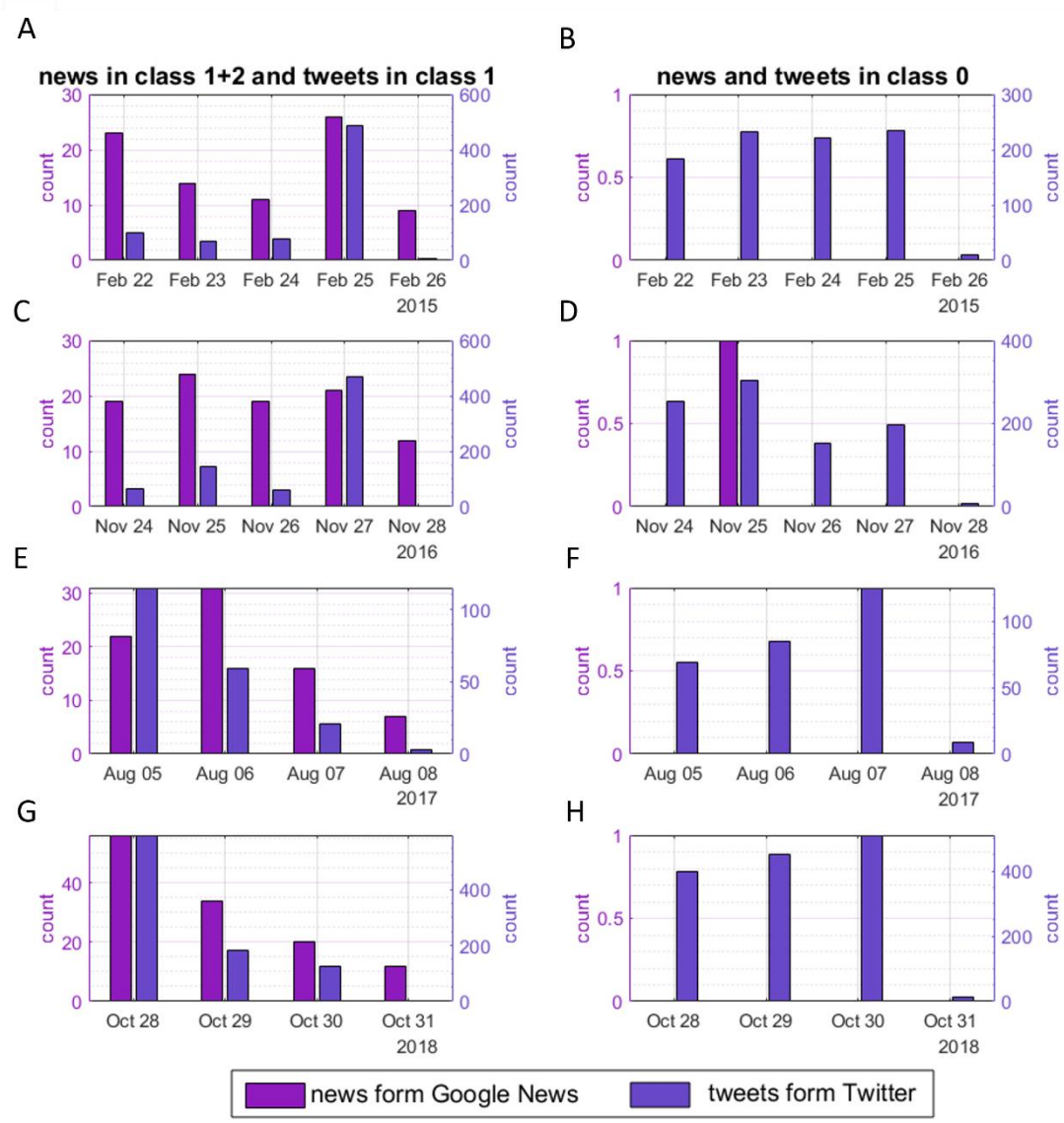

Figure 54: In the first column (**A**, **C**, **E** and **G**) using a bar plot, the time distribution of news in class 1+2 and tweets in class 1 from 2015 to 2018 are featured. The time distribution considers a maximum of 5 days to a minimum of 4 days. In the second column (**B**, **D**, **F** and **H**), the class without information for each variable (news and tweets) has been considered. Panels were generated using MATLAB R2021b.

4.2.2.2 Case study within the tweet dataset

Exploring the tweet dataset, some case studies were considered. The first case examines two landslide events that occurred on the same day of February 25, 2015. They had different localizations and different triggering mechanisms and consequences in terms of economic losses and human lives. The first event is shown in Figure 55A. The Senio River, located in Casola Valsenio in Ravenna Province (Emilia Romagna), with bank erosion undermined the embankment at the base, causing an extensive landslide involving a football field. The landslide involved surface material characterised by arenaceous and arenaceous-marly units from the middle-lower Miocene. No casualties were reported, only infrastructure damage.

On the same day, strong precipitation caused one landslide on the island of Ischia near the coast (Figure 55B). One person was swept away by the said landslide. The material involved was purely superficial and of volcanic origin, characterised by pyroclastic and ignimbrites. The precipitation event was traced within the criticality bulletins issued by the Campania Region. On 24th February 2015 at 13:15, the Campania Region issued a warning for forecasting adverse weather conditions from 16:00 until the following 24 hours (25th February 2015). The weather bulletin described the presence of a perturbation of Atlantic origin forming short showers or thunderstorms locally of moderate intensity.

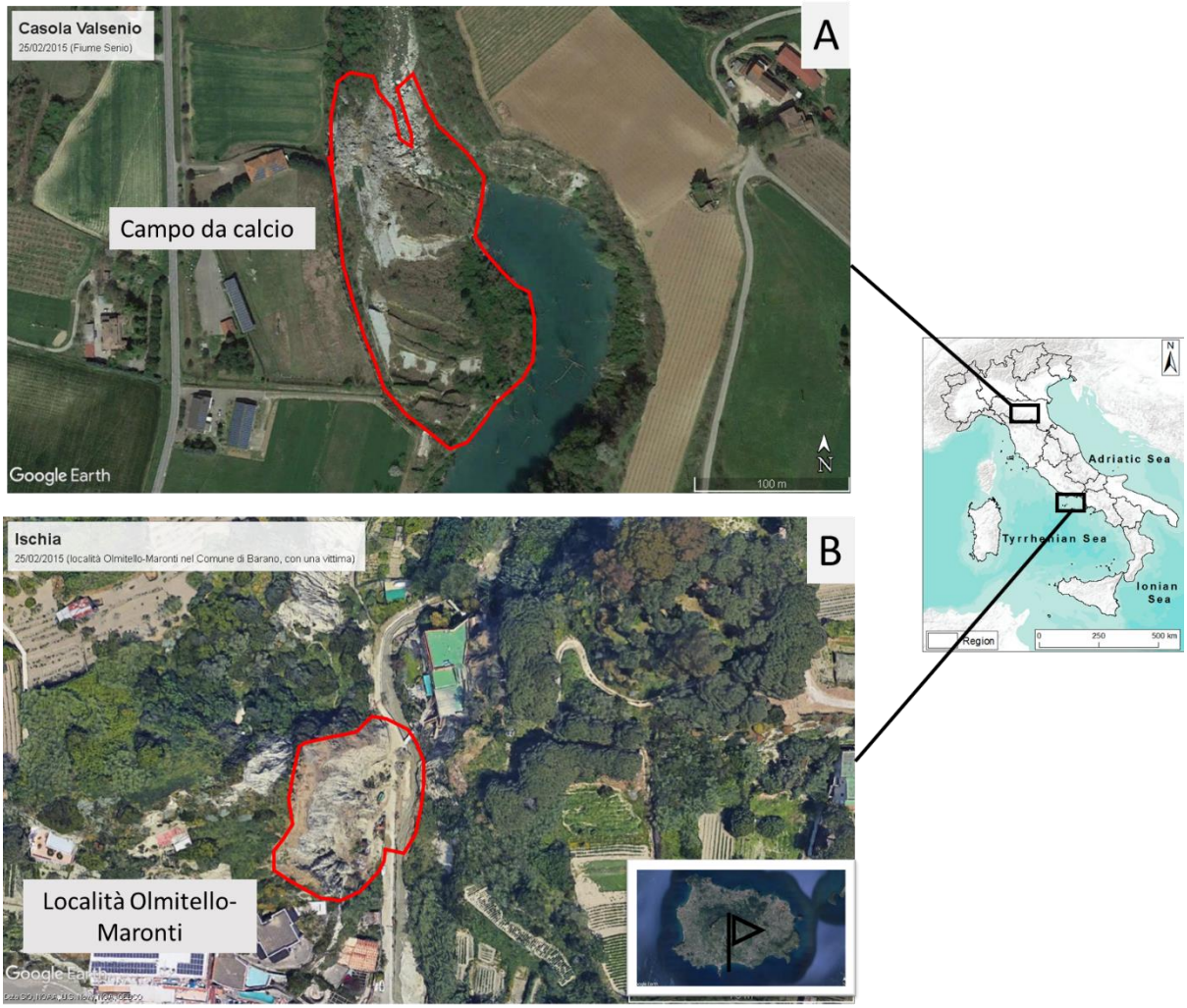

Figure 55: Two case studios, in red is showed the landslide. **A** One landslide was triggered by river erosion in Emilia Romagna at Casola Valsenio on 25/02/2015. **B** On the same day, another landslide was triggered by rainfall at Ischia in Barano municipality. During this event, one victim was reported.

Casola Valsenio in Figure 56A shows homogeneous and limited tweet and retweet distributions and vice versa for the Ischia event shown in Figure 56B. The first tweet described a landslide event in Barano and was posted at 10:45 on February 25, 2015. At 16:02, there was the first tweet with “landslide” and “fatality” as word associations. The Tweets peak was reported at almost 17:00. Eighty-seven tweets were recorded, and 11 were retweets. Then, the record decreased.

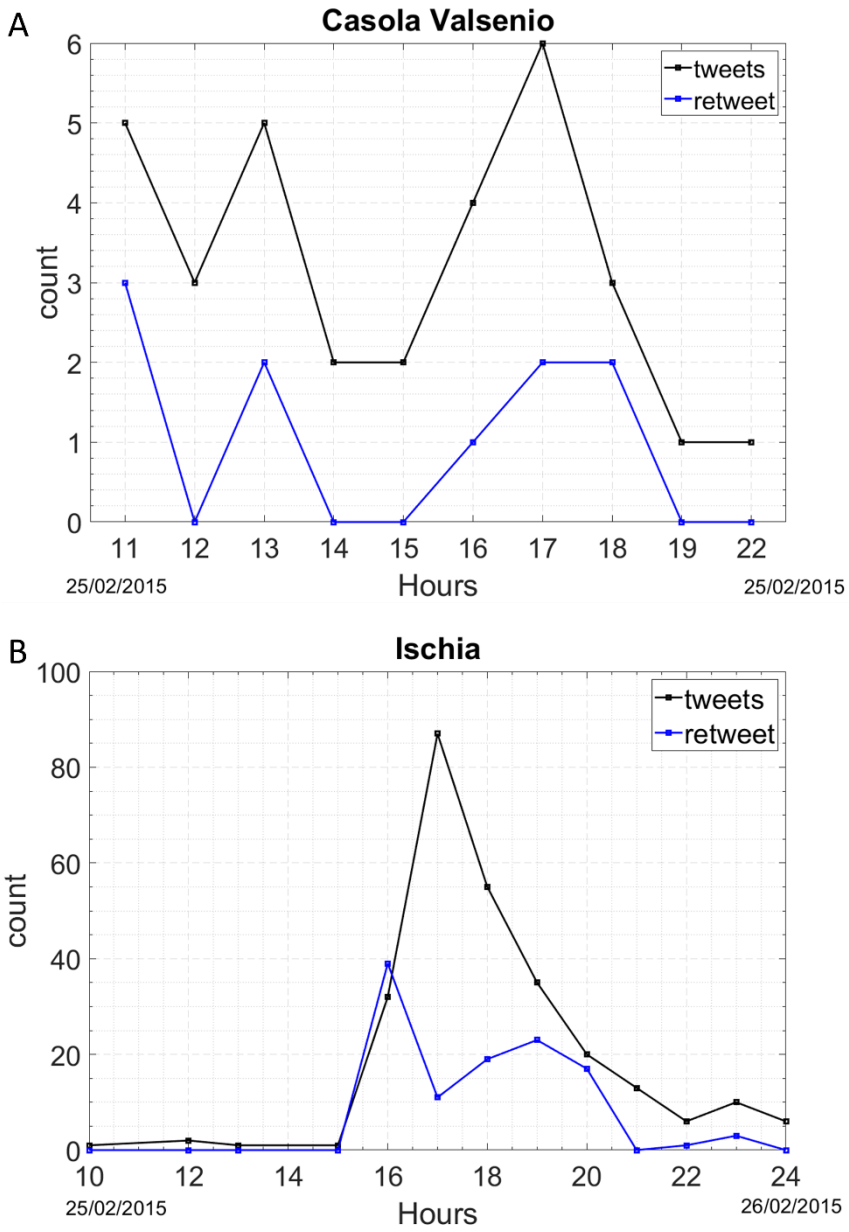

Figure 56: On the same day, two events occurred in Italy and different parts (Emilia Romagna at Casola Valsenio and Campania at Ischia). Panel **A** shows the tweet distribution and their spread with retweets at Casola Valsenio. Panel **B** shows tweets and retweets at Ischia; during this event, one victim occurred. In both panels, the data are shown on an hourly basis. Panels were generated using MATLAB R2021b.

The second study case is shown in Figure 57. The recent event occurred in the afternoon of the 24th of November 2019 in Savona municipality, close to the Madonna del Monte village.

The landslide occurred via the viaduct and included almost 40 metres of the A6 motorway, which links Savona and Torino cities. The landslide involved alluvial and fluvial sediments. The trigger was caused by intensity and extended precipitation on previous days. The crown is located in contact with underlying units lithostratigraphic (mudstones with sandstones-Permian lower) with almost 180 m s.l.m. The landslide was considered a new formation, and the slide was roto-translative. The mobilized material included almost 15 thousand m^3.

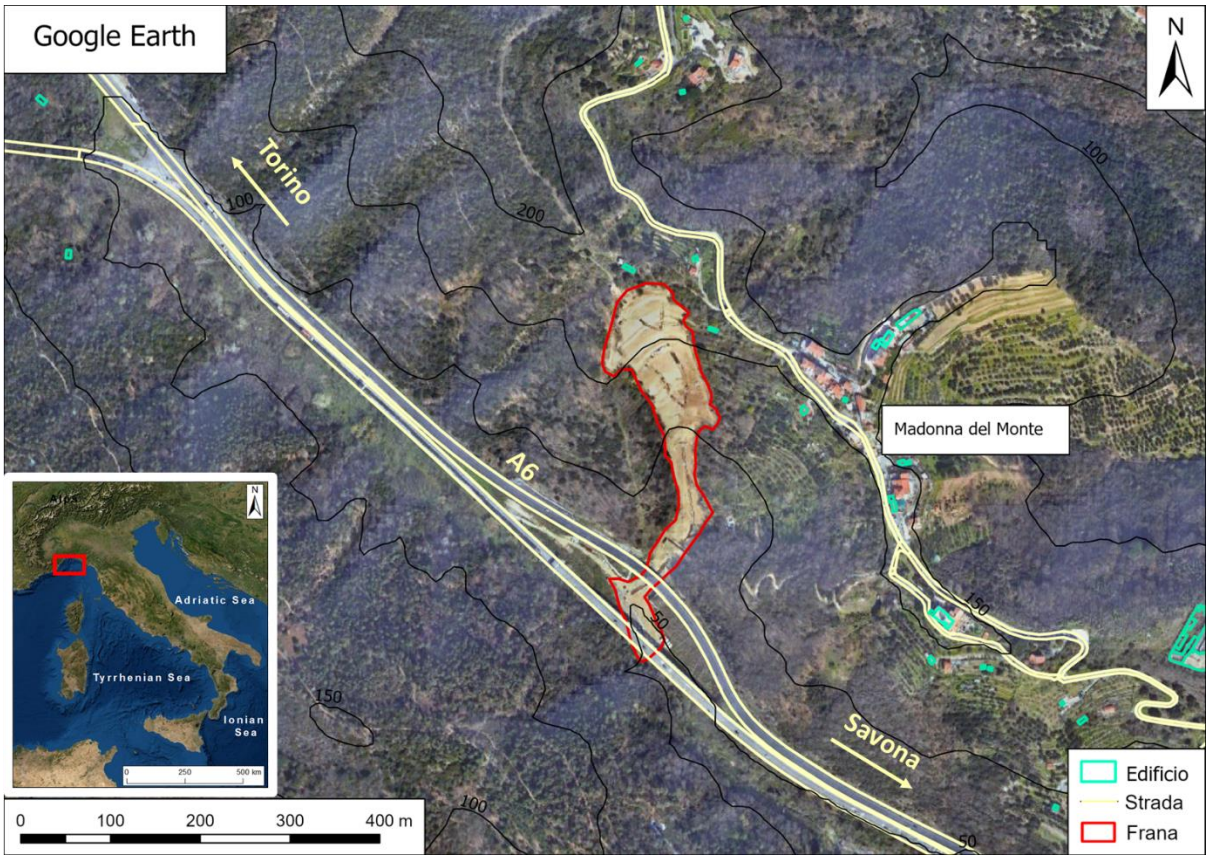

Figure 57: Focus on landslide events in Madonna del Monte near Savona. The landslide occurred via the viaduct of motorway A6 that comes from Savona to Torino.

In the Liguria Region, 128 pluviometers were spread. The landslide near the motorway was triggered by intense rainfall on the previous days on 24th November 2019. For this reason, data from 1st October 2019 to 24th November 2019 were considered for subsequent analyses.

Rainfall data were interpolated from points to create a raster surface using an inverse distance weighted (IDW) technique. The output cell was 300x300.

From 1/10/2019 to 24/11/2019, a cumulative maximum of almost 1850 mm was measured in northern Liguria near the Piemonte region and along the Apennines chain (Figure 58A). Figure 58B shows a focus on the area of landslide events (red flag). Based on the rainfall distribution, the event area did not have significant precipitation values. In fact, a detailed analysis was carried out for each pluviometer with anomaly data. Pluviometers without data feature white anomalies. This aspect is clearer than the four rain gains around the event, causing an underestimation of rainfall distribution in the IDW interpretation.

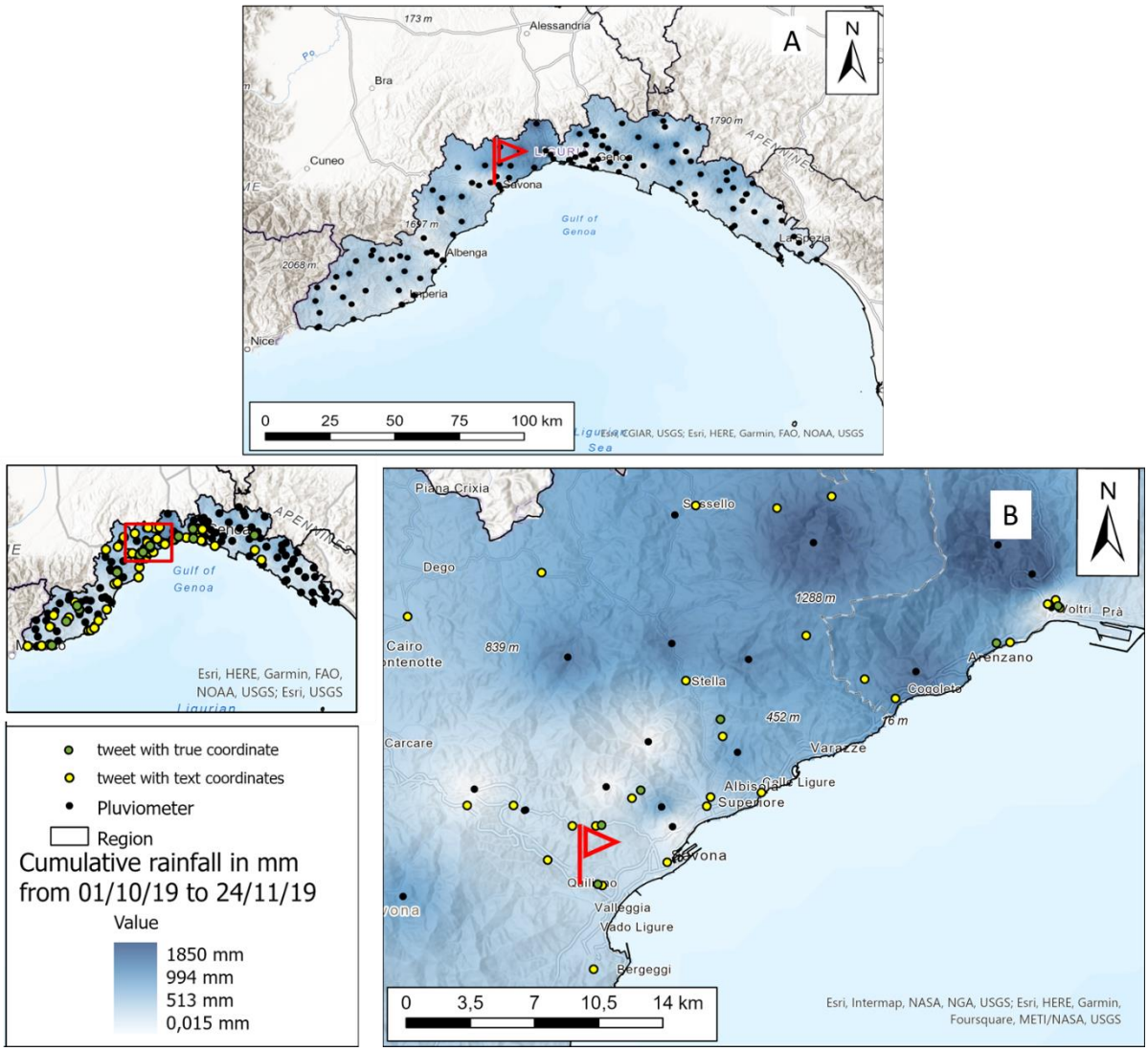

Figure 58: A and **B** include the average rainfall from 1/10/2019 to 24/11/2019 in Liguria. The red flag indicates the landslide event near Madonna del Monte village, east of Savona. The maps were generated using ESRI ArcGISPro.

Moreover, the temporal distribution of rainfall has been analysed at the regional scale. On the basis of rain gauges nearer, data from October until 24 November were regarded. Figure 59A and B display the daily mean and cumulative mean daily. Figure 59A shows data for two months, while Figure 59B illustrates a focus on November. From the two panels, it is clear that the sizeable distribution of precipitation occurred before the 24th. The highest daily mean measured on 23rd November was almost 84 mm, with a cumulative value of approximately 492 mm.

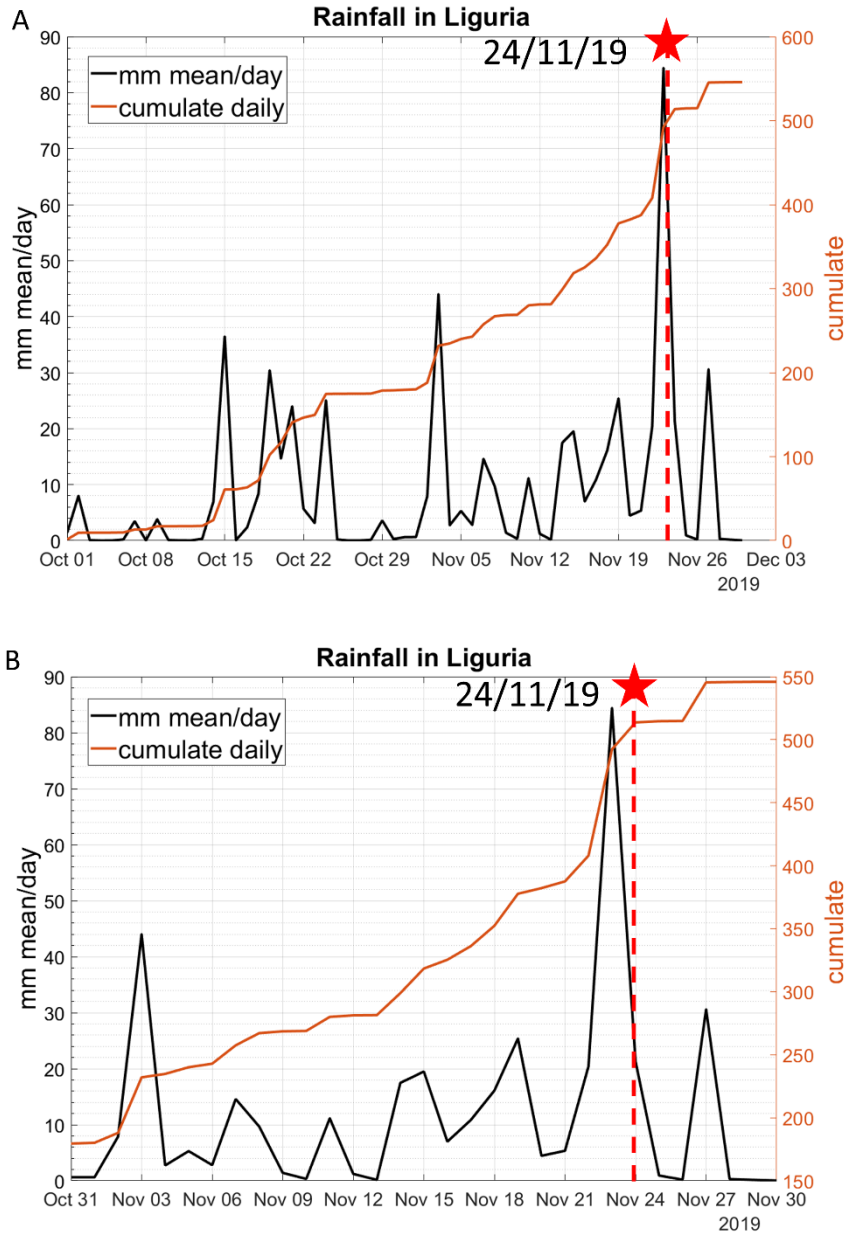

Figure 59: A and **B** rainfall data distributions with mean daily and cumulated for almost two months in Liguria with a focus on November. The point of the event has been signed with a red star and line. Panels were generated using MATLAB R2021b.

Google News has been the second source of information about events to outline the event dimension in terms of speed publication and media impact. On the basis of reports, the first article was published at 14:37, 19 minutes after the event. Figure 60 shows the first article published by <<Il Messaggero>>, and the last article by <<Lo Spiffero>>. This last article was released at 21:29 on 24th November 2019.

Figure 60: Two types of articles within Google News were collected. At 14:37 <<Il Messaggero>> and then at 21:29 <<Lo Spiffero>> published one article about viaduct on A6 with some detail, for example: localisation, victims, type event, damages, and data.

Figure 61A and B show the news distribution during November in Italy and in the Liguria Region. In both cases, the peaks of "Newspaper articles" and "Landslide events" are illustrated during the 25th November. For the 24th of November, in the whole Italian territory, 66 "Landslide events" were harvested and distributed in 89 "Newspaper articles". Of these, 31 "Landslide events" were reported in Liguria in 47 "Newspaper articles" (Figure 61B).

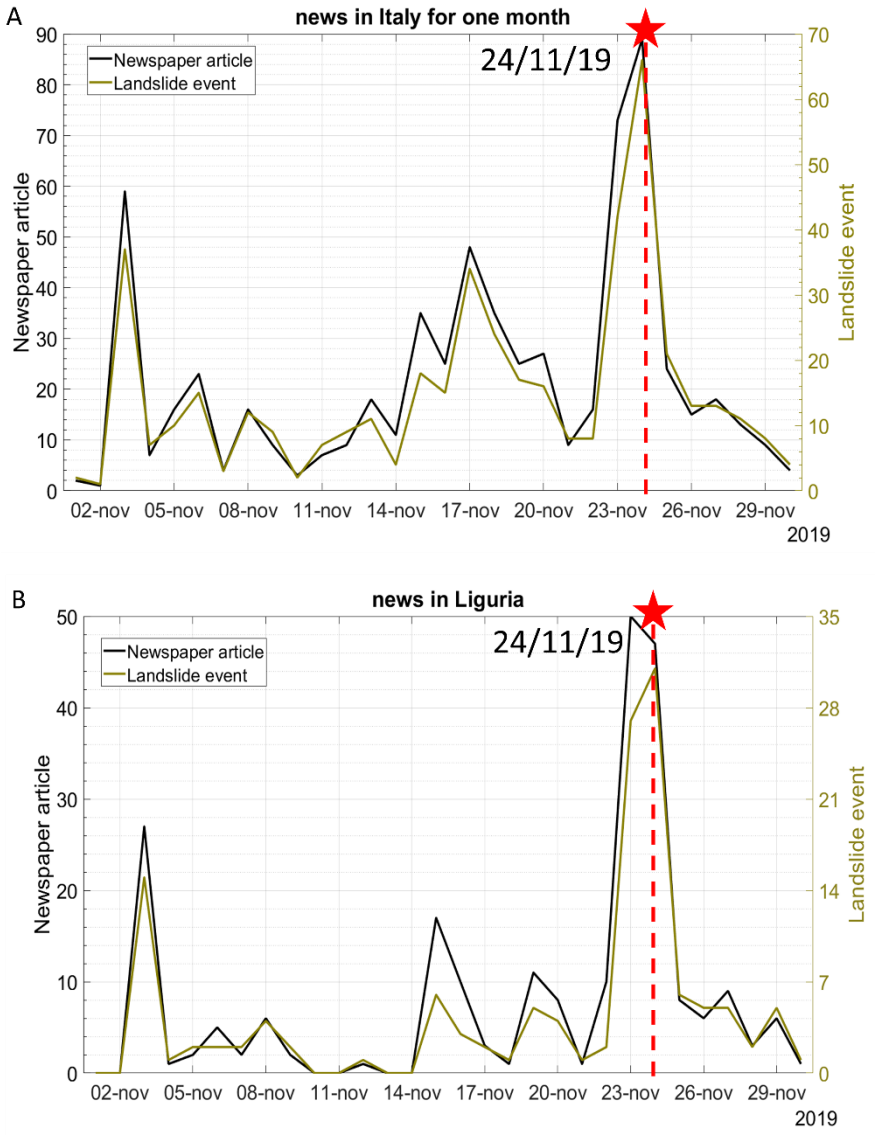

Figure 61*:* In **A** and **B,** the distribution of News inside Google News has been demonstrated. In fact, two distributions have been considered: in general, in the whole Italian territory and then only for the Liguria region. The point of the event has been signed with a red star and line. Panels were generated using MATLAB R2021b.

Using data mining, 2100 tweets during the 24th of November were harvested. Figure 62 presents some tweet examples that describe the landslide event. Tweets from different accounts have been collected, for example, by official channels, such as Regione Liguria, but also from official newsletters (as <<IlGiornale>>) and citizens or amatorial citizens (such as <<Rete Meteo Amatori>>). To a certain extent, tweets provide good information details (such as by users), while others publish photos or videos along with the text.

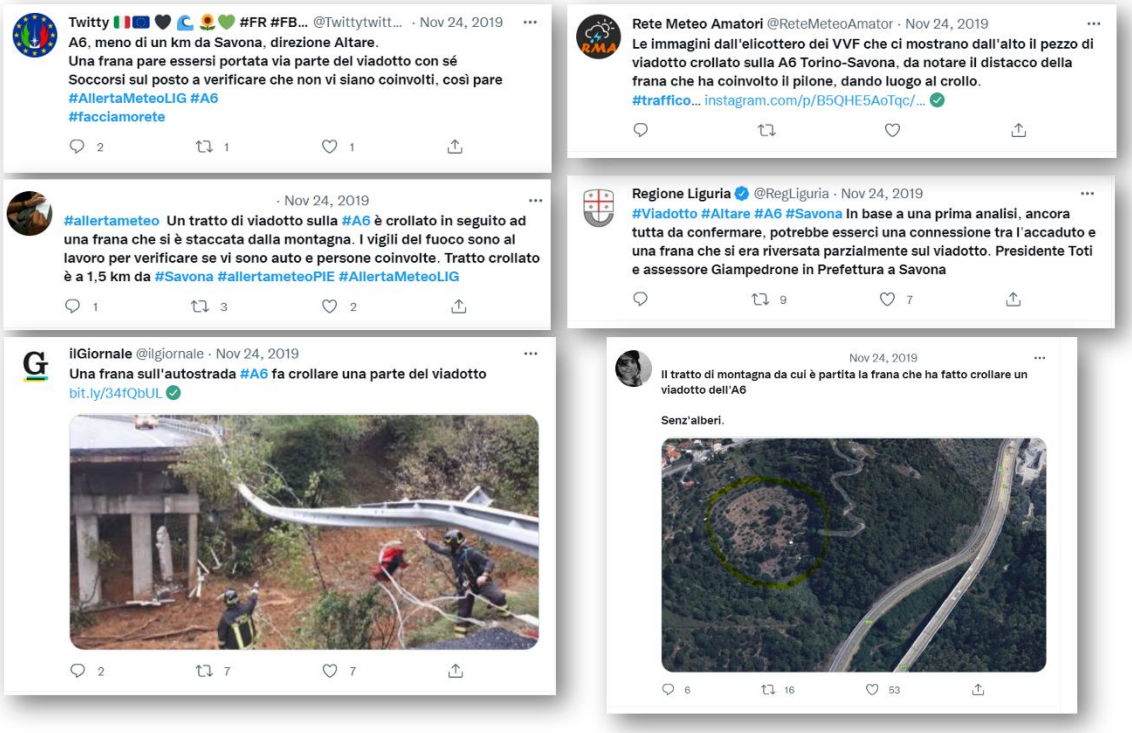

Figure 62: Some examples of Tweets about landslide events on the A6 motorway. Different details about events are possible to obtain, for example, place, time, feasible victims, damages, photos or videos. Several types of users can post specifics of the event: official channel (as Regione Liguria), citizens (Rete Meteo Amatori) or newsletters (as ilGiornale).

It is possible to outline a sequence of events, analysing the dataset:

1. At 14:26, the first tweet that describes the fallen viaduct has been published;

2. At 14:36, the landslide-specific tweet (10 minutes later) has been released;

From this dataset, two panels have been compared, considering general tweets in the whole Italian territory (Figure 63A) and tweets with information about events (Figure 63B). Overall, in Italy on 24 November 2019, 856 of 2100 tweets described landslide events. Retweets were almost 1776. Considering only landslide events in the A6 motorway, 390 tweets were posted with retweets or a media impact of 1061. The peak was measured at 15:00. The publishing tweets continued in the next hours but with a decreasing trend. However, this distribution was considered not to be exhaustive.

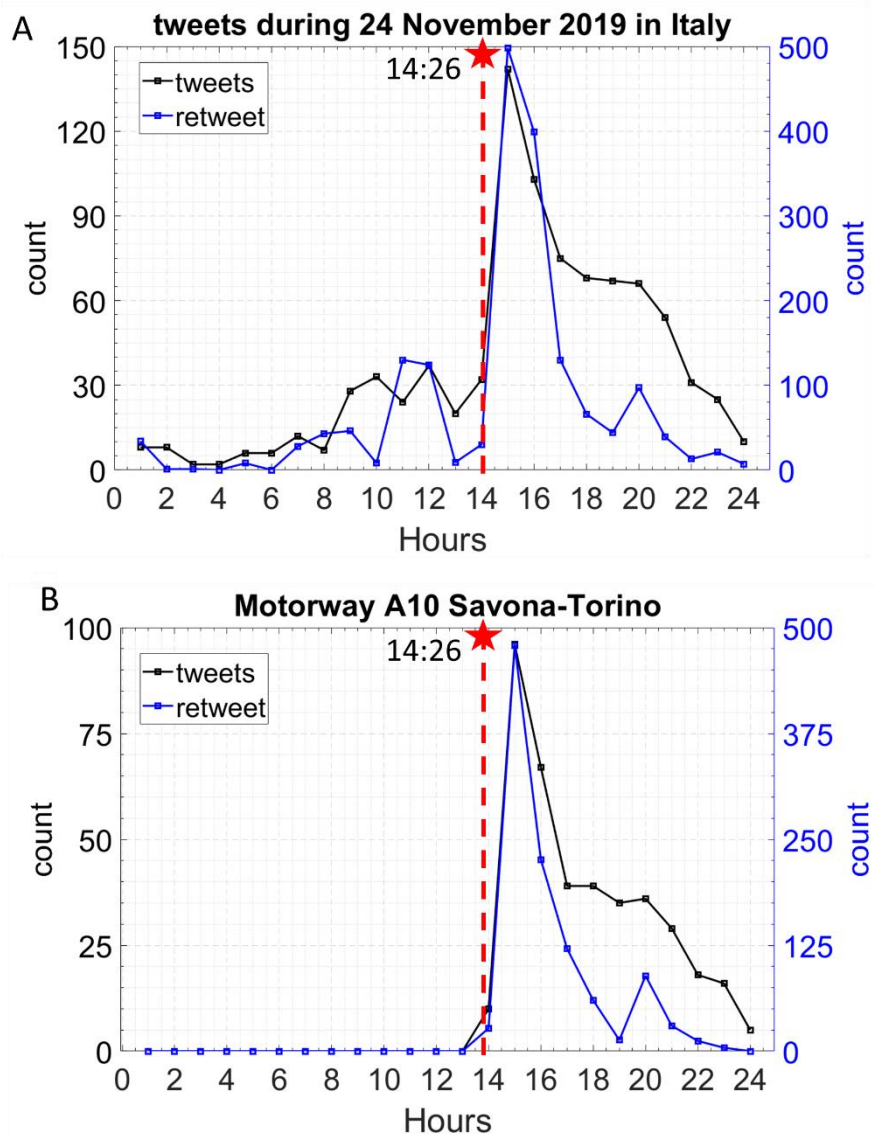

Figure 63: In **A** and **B,** data from Twitter are shown in two drafts: considering all tweets in Italy (**A**) and only tweets that describe the landslide event near motorway A6 (**B**). Both panels consider only 24 November 2019. Panels were generated using MATLAB R2021b.

To outline a possible spatial distribution, the coordinates have been attributed to some data. Two types of coordinates were outlined from the text and effectiveness of the event through photointerpretation. Inverse distance weight (IDW) was applied with a resolution of 307x307 to obtain data homogeneity for "tweet with text coordinates" and 300x300 (Figure 64A and B) for "tweet true coordinates" (Figure 64C and D). Parameters as exponent of distance and the number of points were maintained constant, equal to 2 and 10 respectively. The data used to outline the echo media is the counting of published tweets. Figure 64A and B show a possible spatial distribution using coordinates from the text (yellow point and red flag sign the landslide in the exam). The peak in this case is recorded to North-Est of Savona. The data are coherent with the rainfall distribution. Figure 64C and D the true

coordinates of events are reported in green. In this case, on the basis of tweet counting, important media impact can be measured near Quillano (SV) with possible implications in closed areas.

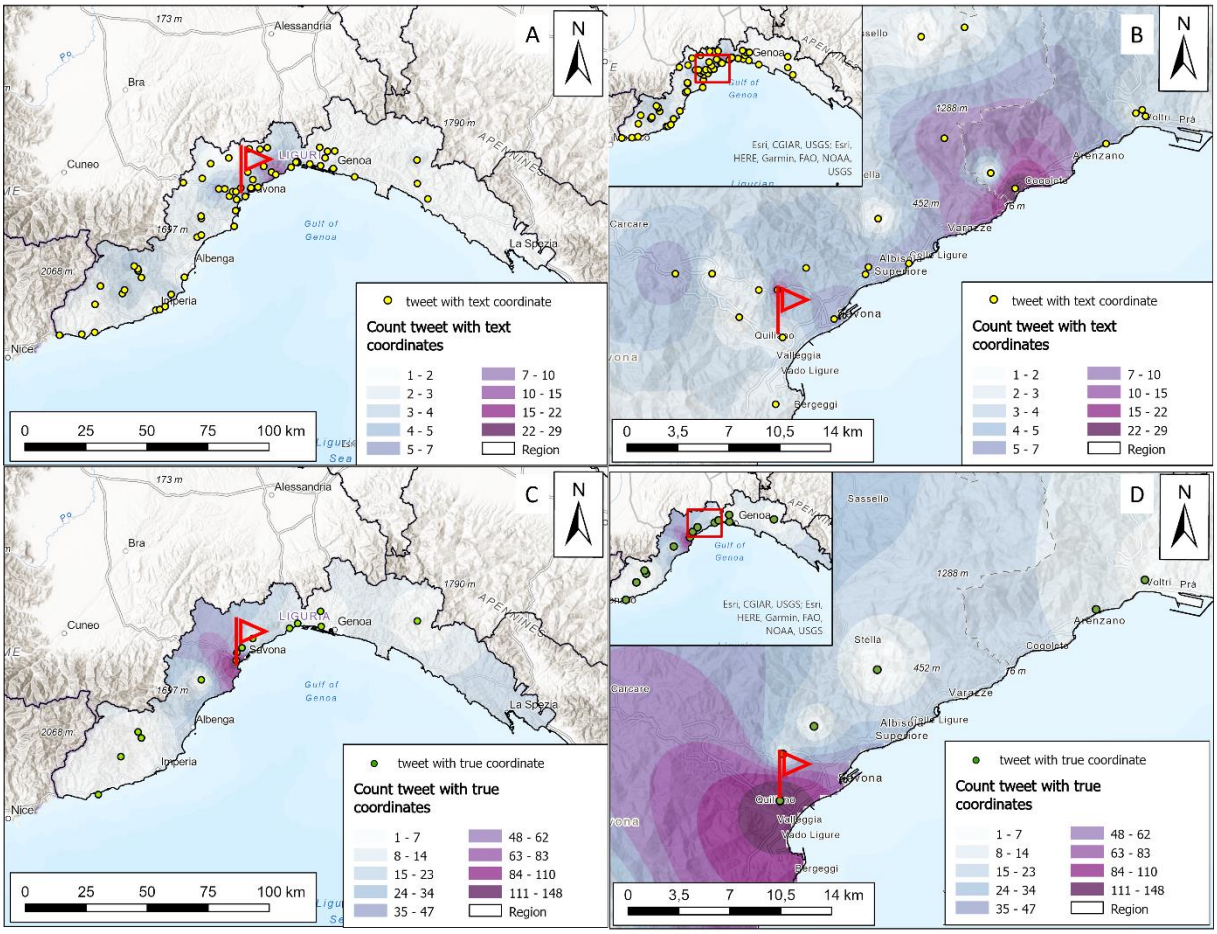

Figure 64: In **A** and **B,** tweets with coordinates from tweet text for the Liguria region have been mapped. From the data points of tweets, an IDW was applied to obtain the spatial distribution and, hence, the media impact of the event. Tweet data referred to only 24/11/2019. Since the tweet distribution to the northwest of Savona, there was a peak of tweets. This is in contrast with **C** and **D,** where tweets with true coordinates have been mapped. In Quillano, the epicentre and, hence, maximum media impact of the event were reported. The maps were generated using ESRI ArcGISPro.

From counting tweets, it is possible to obtain a map with a possible alert system. Figure 65 shows different analyses on the basis of the coordinates utilised. Figure 65A and B display municipalities and Alert Zones in the Liguria region considering the coordinates from tweet text. The highest values are localised in the central region in closed municipalities around Savona and Genova. It is noted that tweets do not cover all municipalities in Liguria. Actually, the tweet volume for some counties was pretty low, possibly because of a smaller population or limited interactions by users about events.

Two alert zones (Figure 65B) present sizeable values of tweet count and are located in the central and western regions. Only one does not show significant counting, placed in the northern area.

Figure 65C and D present two possible maps using true coordinates of tweets, to municipality and alert area scale. Figure 65C shows localized important values of counting in Genova and Savona. The alert areas with significant values were concentrated along the sea (Figure 65D). Two alert zones in the northern region present low values in contrast with the sizeable rainfall measure (Figure 58A).

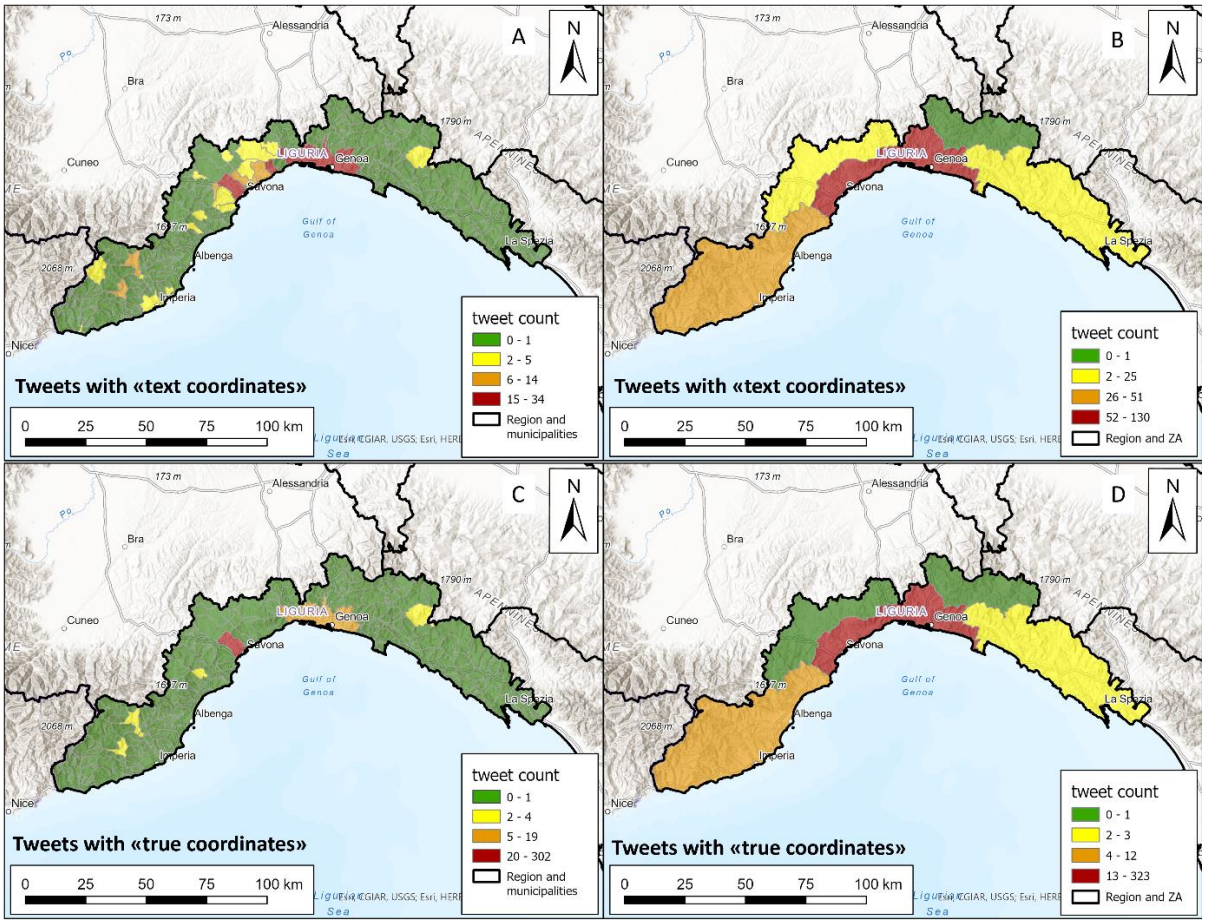

Figure 65: Tweet count based on tweets with text coordinates and true coordinates. **A** Presents the municipality scale of the distribution of tweets. Significant values are localized in the central region in closed municipalities around Savona and Genova. **B** The distribution is shown to alert areas with only one not showing significant counting. **C** Displays the tweets counting using the true coordinates of the event on a municipality scale. Savona and Genova present sharp values of tweet count. **D** Considering alert areas, with two alert zones with low tweet counts, alert areas with significant values are concentrated along the sea. The maps were generated using ESRI ArcGISPro.

Applicability of the data

During the classification of the Twitter dataset and subsequent analysis on some relevant landslide events, obvious classification difficulties were also noted. The Twitter dataset was found to be significantly noisy and uneven in the information provided with respect to the event on-going. For example, the phrase:

"Collapsed portion of viaduct on A6 Turin-SavonaWe are officially in total and isolated disaster. #A6 #collapse #Liguria @vlp31 @Agency_Ansa" (translation: "*Crollata porzione di viadotto sulla A6 Torino-SavonaSiamo ufficialmente nel disastro totale e isolati. #A6 #crollo #Liguria @vlp31 @Agenzia_Ansa*").

It cannot be classified as a natural event. From this example, it is clear that the word "landslide" ("frana") needs to be accompanied by an additional specification. The word "landslide" ("frana") or "collapse" ("crollo") especially in the Italian language is used for different contexts and the specification avoids creating ambiguity with respect to the event considered.

Changing the meaning to: "Collapsed portion of viaduct on A6 Turin-Savona *due to landslide*. We are officially in total and isolated disaster. #A6 #slide #Liguria @vlp31 @Agency_Ansa" (adding: *causato da una frana o dovuto da una frana*).

This can actually be considered a hydrogeological hazard event. Furthermore, textual data are very lacking in information if the user is simply a witness or an ordinary citizen, while more information can be extracted from official channels, such as Fire Department or the Region or amateurs in the field. Such aspect can be an advantage, but at the same time, it creates a lack of uniformity in the language used by rescue managers. It is the most important and significant challenge to create a homogeneous language at least between official channels or organs of Civil Protection. Such cooperation can available communication between decision-makers and citizens, but also decision-makers and data analysis-makers. Such considerations can bring contribution to the implementation of specific communication and warning guidelines with respect to natural events, such as landslide hazards, as in this work discussed. A possible example is to create a text with the below specifics:

- Event entity: "frana terreno", "frana in roccia", "frana scivola" (same example with the word "smottamento"), "crolla terreno", "crollo in roccia", "crollo in roccia staccatosi"; in some cases using appropriate articles "frana lungo la strada", "frana nella strada", "frana sulla strada", "crollo in roccia lungo la strada", "crollo di rocce nella strada", "crollo di rocce sulla strada", "strada coinvolta in crollo di roccia"; while with other terms have to be highlighed further specifics "frana porzione di terreno", "strada scivola causa frana".
- Place: where is the event; it should be specified in three possible manners, 1) using words with hierarchy sequence, municipality, provincial, regions; 2) geolocation of the event with coordinates provided by users; 3) geo-localization of users.
- Other information: victims, damage, rescues, sentiment, opinion, photo, video etc.
- Time can be withheld because it is possible to get through the entities of tweet during extraction.

In particular, the coordinates extracted from the text can be useful for creating appropriate hazard maps, which are useful to the national coordination centre during the task of monitoring and surveillance of the national territory in order to identify planned or ongoing emergency situations and follow their evolution, as well as to alert and activate the various components and operational structures of the National Service of Civil Protection that contribute to the management of emergencies. In fact, considering the current warning emergency about weather events it is possible to implement these maps considering the spatial distribution of tweets publications or news publications.

Below is shown an example of possible text with all necessary's information on ongoing landslide event:

"Landslide along SP49 road near the car park of the cemetery in the municipality of Sestino in the province of Arezzo (Tuscany). The slide caused damage. There were no victims. Operators are already on their way to the area for restoration. #landslide #Tuscany".

Translation: *"Frana lungo la strada SP49 all'altezza del parcheggio del cimitero nel comune di Sestino in provincia di Arezzo (Toscana). Lo scivolamento ha causato danni. Non ci sono state vittime. Gli operatori stanno già raggiungendo l'area per il ripristino. #frana #Toscana".*

4.3 Applying BERT

The manually classified dataset provides a solid base for applying deep learning using the natural language technique. The dataset was utilised as an anchor point for supervised learning of the deep learning method to classify tweets. A script was created for text classification, and it is capable of distinguishing whether a tweet describes a landslide event. The script is henceforth called "**Bert For Information on Landslide Events**" or BEFILE. BEFILE trained on the aforementioned classified dataset in the Italian language.

Three types of preprocessing were applied to obtain the best results with XLM-RoBERTa as the model. The first preprocessing considered the dataset without cleaning; the second considered all possible parameters of removing. Finally, in the third proposal, only some parameters have been removed from the text.

Subsequently, the dataset from the newspaper has been correlated to the classified dataset by BIFILE. This process allowed for the validation of the classified tweets dataset. Different strategies have been utilised, considering three types of results. Temporal distribution was analysed for each target (0: "No

Landslide" and 1: "Landslide") and the correlation grade was calculated using nonparametric systems. Such action allowed for the outline of the best model.

Twitter during extraction provided further parameters, such as entities, which provide metadata and additional contextual information about content posted on Twitter. The entities section includes an array of common things within Tweets: hashtags, user mentions, links, stock tickers (symbols), Twitter polls, attached media, location with coordinates, etc. Through entities within tweets, a simple analysis has been carried out to find a solution for coordinates.

4.3.1 Text classifications with deep learning

Three tests of BEFILE have been carried out to obtain text classification, changing the setup of pre-processing. Each model showed the loss and accuracy trends of the training and validation datasets. Furthermore, each model presents one report with main metrics. Graphically, a confusion matrix and receiver operating characteristic with area under the curve were utilised.

The datasets were randomly divided into 80% training and 20% testing. The training dataset features 10.679 data spread in 6850 with target 0 (“No Landslide”) and 3829 data with target 1 (“Landslide”). This dataset was further randomly divided by 20%, resulting in the validation dataset. This operation was carried out for each of the three tests. Only the test set was kept constant. The test set is characterized by 2670 data, spread in 1694 with target 0 (“No Landslide”) and 976 with target 1 (“Landslide”).

The first model was named “Model without preprocessing” because the text was not cleaned or preprocessing was applied. Figure 66A and B show the trend of loss and accuracy of the training and validation tests during the training of the model. The model finished 70 epochs because there was no improvement after 15 steps (EarlyStopping). Each epoch lasted almost 29 minutes and 40 seconds. Overall, 35 hours (1 day and half) were required to train and validate the model.

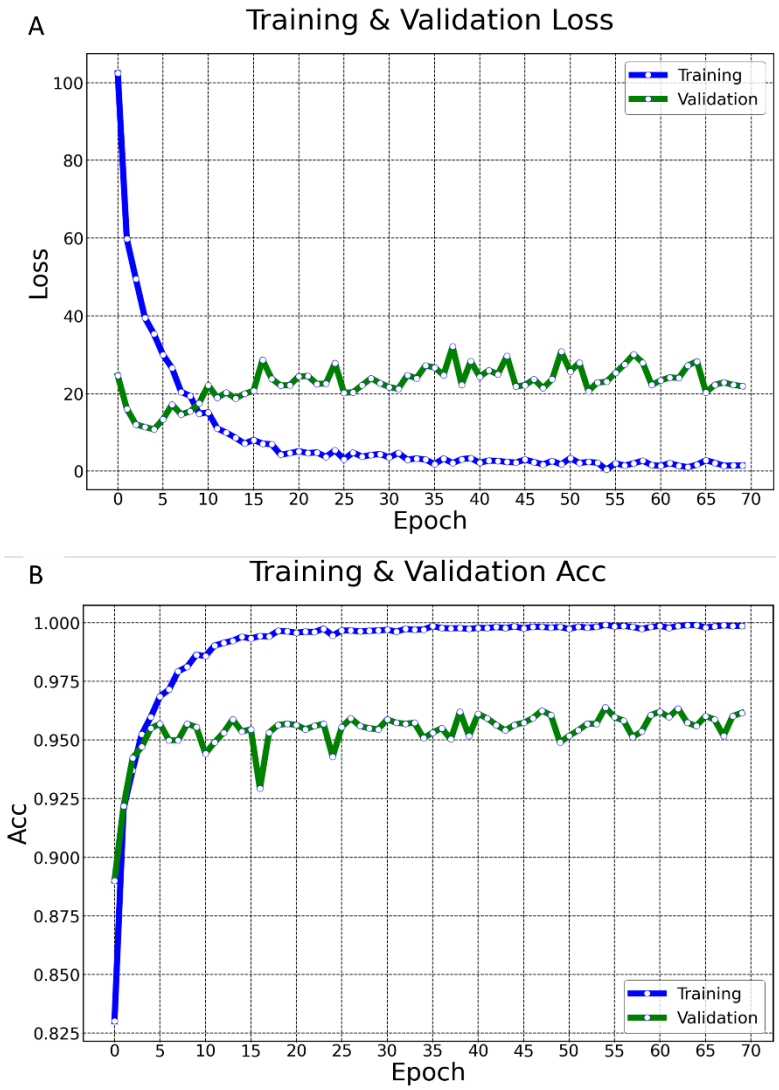

Figure 66: Trend during model training without preprocessing. **A** Panel shows the Training and Validation Loss trend. **B** Panel with Training Validation Accuracy trend. Panels were generated using Python.

The model achieved an accuracy maximum of 0,96. Subsequently, it was tested on the test set. The results of the testing have been reported using metrics in Table 16. Significant parameters of precision, recall and F1 score were recorded for target 0 (coherent with the high distribution of this class).

Target	Precision	Recall	F1 score	Support
0	0,96	0,97	0,97	1694
1	0,94	0,94	0,94	976
Accuracy			0,96	2670
Macro	0,95	0,95	0,95	2670
weighted	0,96	0,96	0,96	2670

Table 16: Metrics used to obtain the accuracy of the model. For each target, the precision, recall and F1 score were calculated. For the model without preprocessing, an accuracy of 0,96 has been reported. Significant F1 scores for each target were recorded.

Figure 67A reveals the confusion matrix. In total, 2553 of 2670 data points are between “No Landslide” (dark blue in Figure 67A) and “Landslide (light blue in Figure 67A). Only 62 False-Positives and 55 True Negatives (white in Figure 67A) were harvested. Figure 67B describes the ROC curve with an area under the curve (AUC) of almost 0,95.

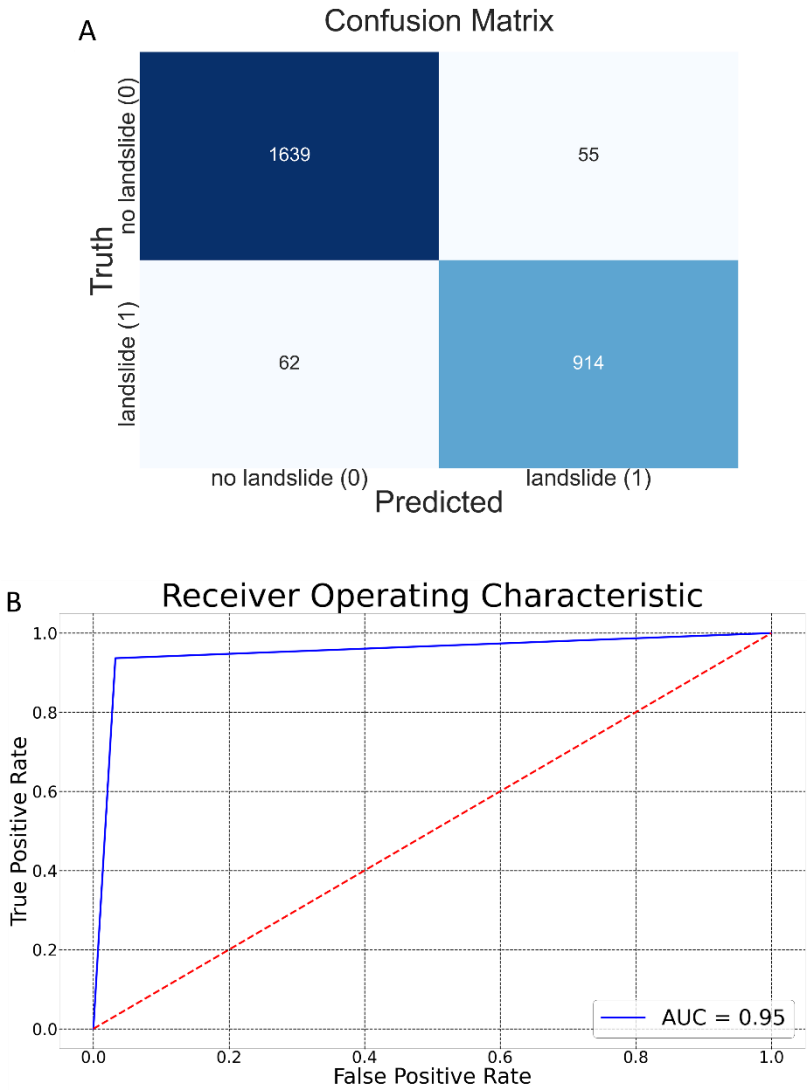

Figure 67: After training, the model was tested on the classified dataset. The confusion matrix in **A** and the ROC curve and AUC in **B.** Panels were generated using Python.

The second modelling was iterated with extreme cleaning text. During preprocessing, several parameters were removed or modified:

- Remove: HTLM special entities, Italian stop words, tickers, numbers, hyperlinks, hashtags, punctuation, special characters, words with 2 or fewer letters, whitespace including new line characters, single space remaining at the front of the tweet and emoticons.
- Modified: @username to AT_USER, lowercase.

The BEFILE model was named “Model with extreme preprocessing”. The epochs were 30. Each epoch lasted almost 29 minutes for a total of 15 hours (half day). Figure 68A and B illustrate the loss and accuracy of training and validation. Compared with the preview model, in this case, the loss is superior, and the time of training is reduced by half.

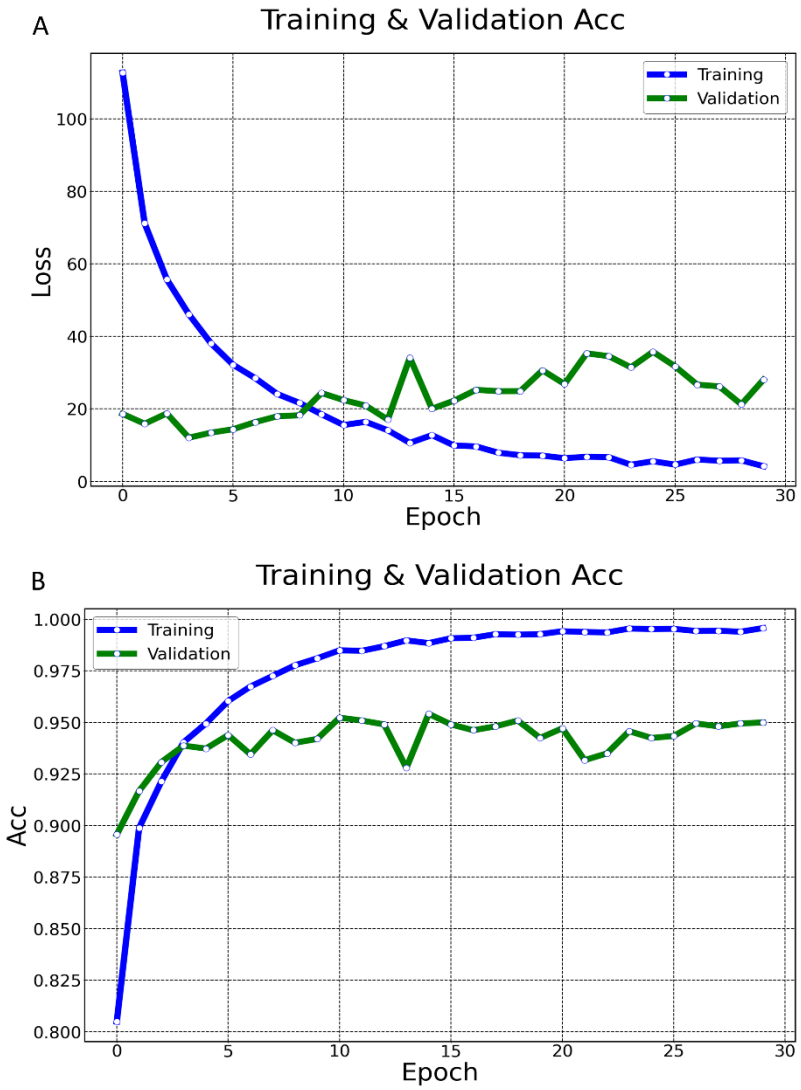

Figure 68: Trend during model training with extreme preprocessing. **A** Panel shows the Training and Validation Loss trend. **B** Panel Training Validation Accuracy trend. Panels were generated using Python.

Table 17 illustrates the metrics. In general, the accuracy reached a maximum of 94%. The F1 score reported significant values for target 0 with 0,95 and for target 1 with 0.91.

Target	Precision	Recall	F1 score	Support
0	0,94	0,96	0,95	1694
1	0,92	0,90	0,91	976
Accuracy			0,94	2670
Macro	0,93	0,93	0,93	2670
weighted	0,94	0,94	0,94	2670

Table 17: Metrics used to obtain the accuracy of the model. For each target, the precision, recall and F1 score were calculated. For the model with extreme preprocessing, an accuracy of 0.96 has been reported. Good F1 scores for each target were recorded.

Figure 69A shows the confusion matrix. Overall, 2503 of 2670 data points have been reported correctly, including 1622 “no landslides” (dark blue in Figure 69A) and 877 “landslides” (light blue in Figure 69A). 99 False-Positives and 72 True Negatives have been archived (white in Figure 69A). Figure 69B describes the ROC curve with an AUC of 0,93.

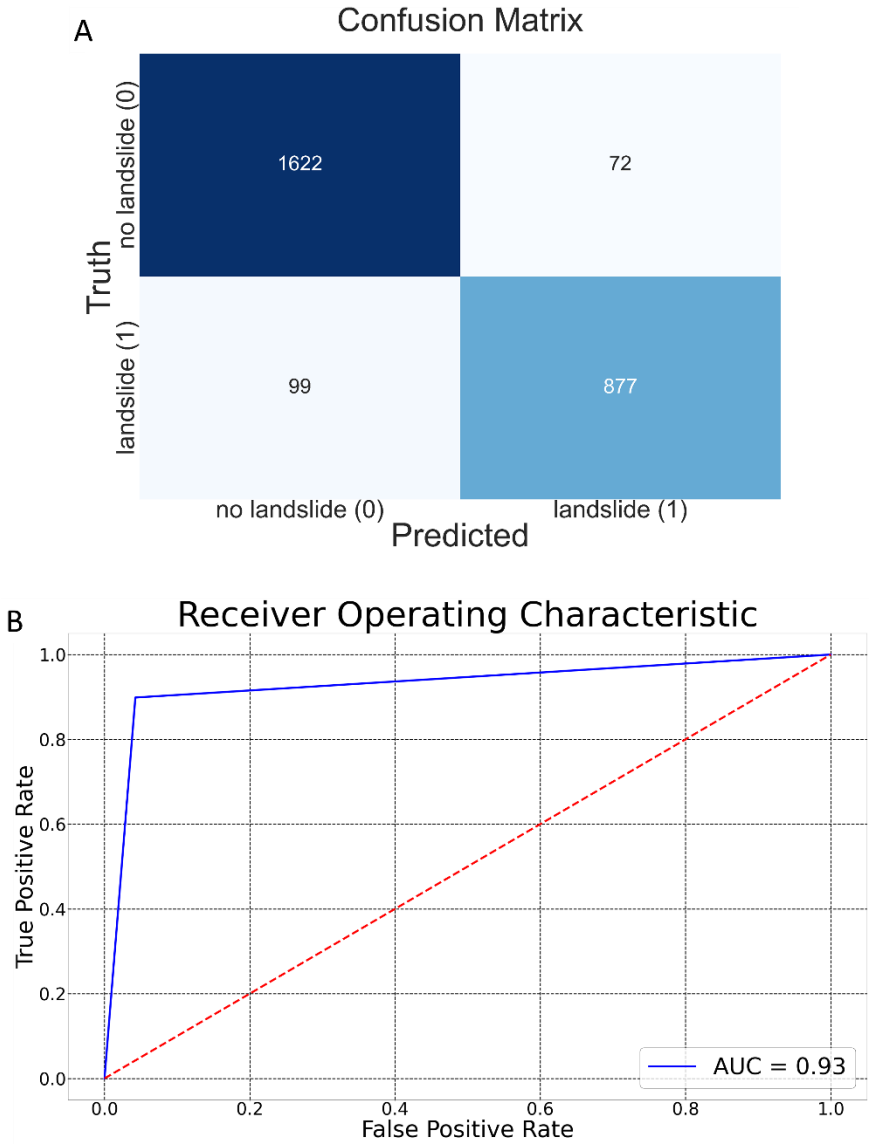

Figure 69: After training, the model was tested on the classified dataset. The confusion matrix in the **A** and Roc curves and the AUC in **B.** Panels were generated using Python.

Finally, modelling was applied with low and random preprocessing. Only some parameters have been removed or modified from the tweet text:

- Remove: hyperlinks, whitespace including new line characters, single space remaining at the front of the tweet and emoticons.
- Modified: lowercase.

Consequently, the model with preview setting data was called the “Model with middle preprocessing”. The training concluded after 42 epochs. Each epoch lasted almost 31 minutes and 4 seconds, for a total of 21 hours (almost one day). With respect to preview models, in this case, the time of iteration for each epoch increases. Figure 70A and B illustrate the loss and accuracy trends.

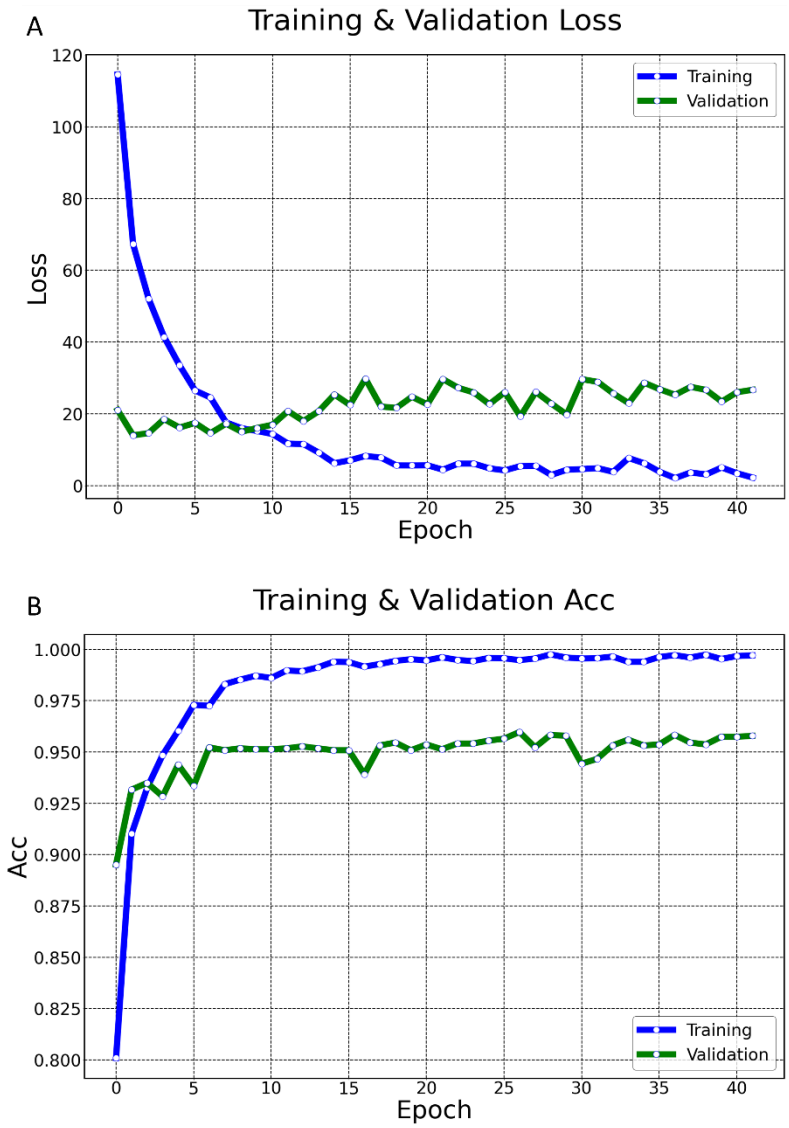

Figure 70: Trend during model training with middle preprocessing. **A** Panel shows the Training and Validation Loss. **B** Panel Training Validation Accuracy trend. Panels were generated using Python.

Table 18 lists the metrics for outlining the performance of the model with middle preprocessing. The model reached an accuracy of 0,95. Important values have been recorded by F1 for both target 0 with 0,96 and target 1 with 0,93. In contrast with the model with extreme preprocessing, the values increased for each considered metric.

Target	Precision	Recall	F1 score	Support
0	0,95	0,97	0,96	1694
1	0,94	0,91	0,93	976
Accuracy			0,95	2670
Macro	0,95	0,94	0,94	2670
weighted	0,95	0,95	0,95	2670

Table 18: Metrics used to obtain the accuracy of the model. For each target, the precision, recall and F1 score were calculated. For the model with extreme preprocessing, an accuracy of 0,96 has been reported. Important values of the F1 score for each target were recorded.

Figure 71A and B display the confusion matrix and ROC curve. A total of 2533 of 2670 data points have been archived with good results (Figure 71A). These are spread in 1641 data in “No Landslide” (dark blue in Figure 71A) and 892 data in “Landslide” (light blue in Figure 71A). Eighty-four False-Positives and 53 True Negatives have been predicted (white in Figure 71A). This iteration showed the best values of False Negative (1641) and as a consequence of True Negative (53), in contrast to previous models. Hence, it succeeds in classifying better tweets without information about landslides (class 0).
The ROC curve presents a good area under the curve of 0,94 (Figure 71B).

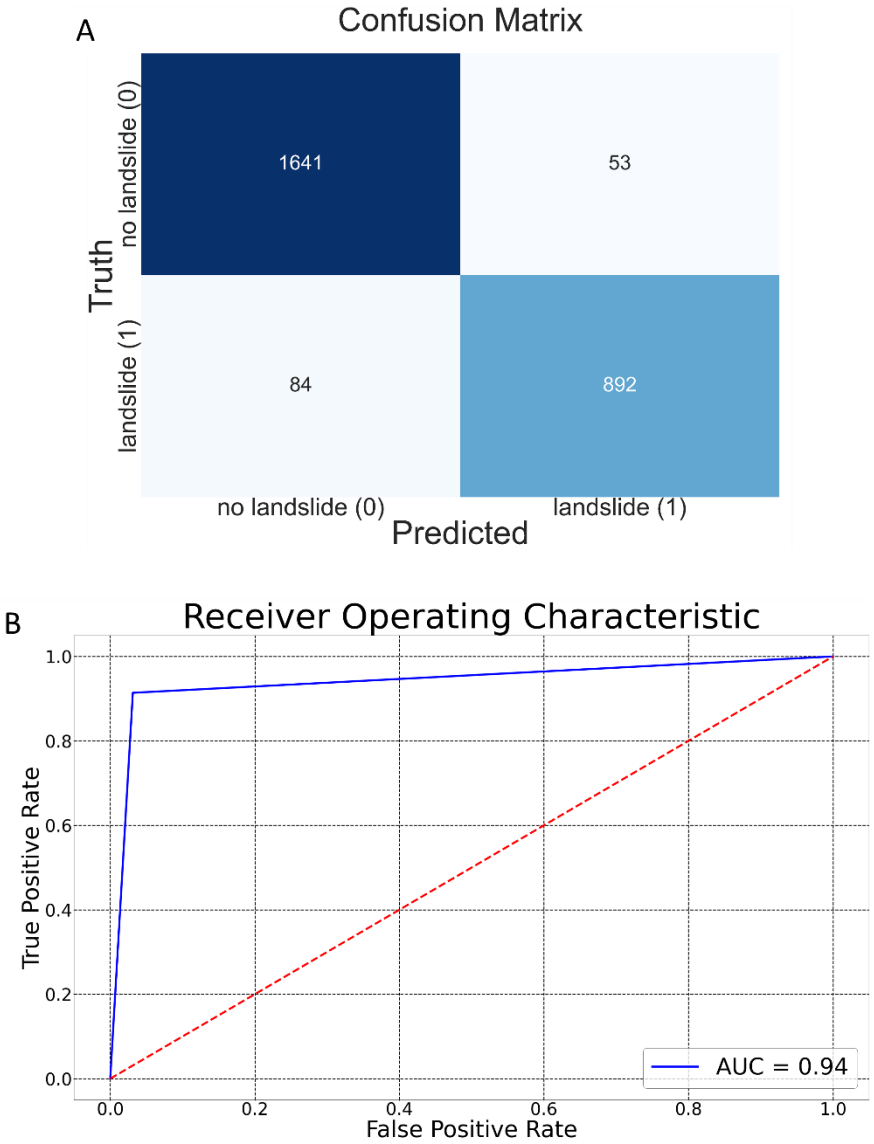

Figure 71: After training, the model was tested on the classified dataset. The confusion matrix in **A** and the receiver operating characteristic curve (ROC) and area under the curve (AUC) in **B.** Panels were generated using Python.

4.3.2 Validation with news dataset

Furthermore, validation was carried out using two case studios. The first study considers a comparison between newspaper articles and extracting a new dataset from Twitter. The second study includes a detailed analysis of the previous case study in the Liguria region.

One dataset from Google News was archived during 2020. The dataset, similar to previous databases, was classified manually into three classes. The classification was based on landslide information, localization and time. The dataset is characterised by 3464 in class 1, 9483 in class 2 and 934 in class 3, for a total of 13.781 data points. Figure 72 displays the distribution of targets during the year. Within

2020, one period was chosen based on important values of article publication. Figure 72 highlights in red rectangle the choosing period for data mining within Twitter. The chosen dataset is utilised to validate the new dataset from Twitter.

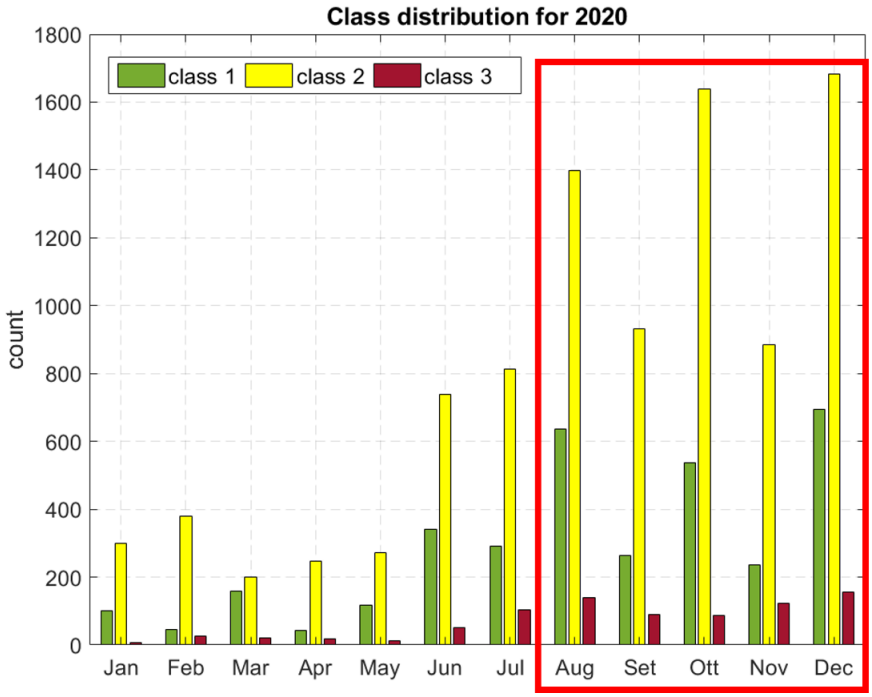

Figure 72: Temporal distribution of news published in Google News for 2020. The chosen period is highlighted in red, from 1st August to 31st December. The panel was generated using MATLAB R2021b.

The dataset from Twitter features 39.780 data points from 1 August 2020 to 31 December 2020.

The validation dataset was submitted to classification using the BEFILE model with the highest accuracy score. The first model of BEFILE without preprocessing achieved the best performance with 96% accuracy. Before applying BEFILE, different sets of cleaning were utilised for the validation dataset. Three cleaning steps were applied. The first case considers the dataset with all text characteristics (without cleaning). In the second case, each arguable interference within the text was removed (HTML special entities, stop Italian words, tickers, numbers, hyperlinks, hashtags, punctualization, special characters, words with 2 or fewer letters, whitespace (including new line characters and single space remaining in front of the tweet were removed). White, @username to AT_USER and lowercase were modified). The third case was removed: hyperlinks, whitespace (including new line characters), and single space remaining at the front of the tweet were removed. The characters have been modified in lowercase.

Each cleaning was named an unclean tweet (preprocessing not applied), clean tweet (extreme preprocessing) or a little clean tweet (some parameters removed) for each model. Table 19 presents the results.

Model type	Target 0	Target 1
BEFILE – unclean tweet	35.787	3993
BEFILE – clean tweet	37.284	2506
BEFILE – little clean tweet	36.339	3441

Table 19: Each model is illustrated quantitatively for each target.

The time of classification is approximately one hour and half for each iteration. The model identified the highest target number in class 1, and hence, tweets with landslide information were the first (BEFILE without preprocessing). The second model with preprocessing identified fewer targets in class 1.

To outline which BEFILE found the best distribution, a correlation with newspaper articles published within Google News was applied. The correlation considered mainly the temporal distribution and not spatial distribution. The temporal distribution of news and each single dataset is shown in Figure 74. Two different sections of news have been analysed: news in class 1 and news in class 1+2.

In the first case, the daily distribution of each variable follows a good trend. (Figure 73) In fact, the peak of posted tweets corresponded to the peak of news. The classification of tweets is not based on the time of the landslide; for this reason, the sum of news class 1+2 has been considered. Similar analysis has been carried out considering even the monthly distribution of data (Figure 74).

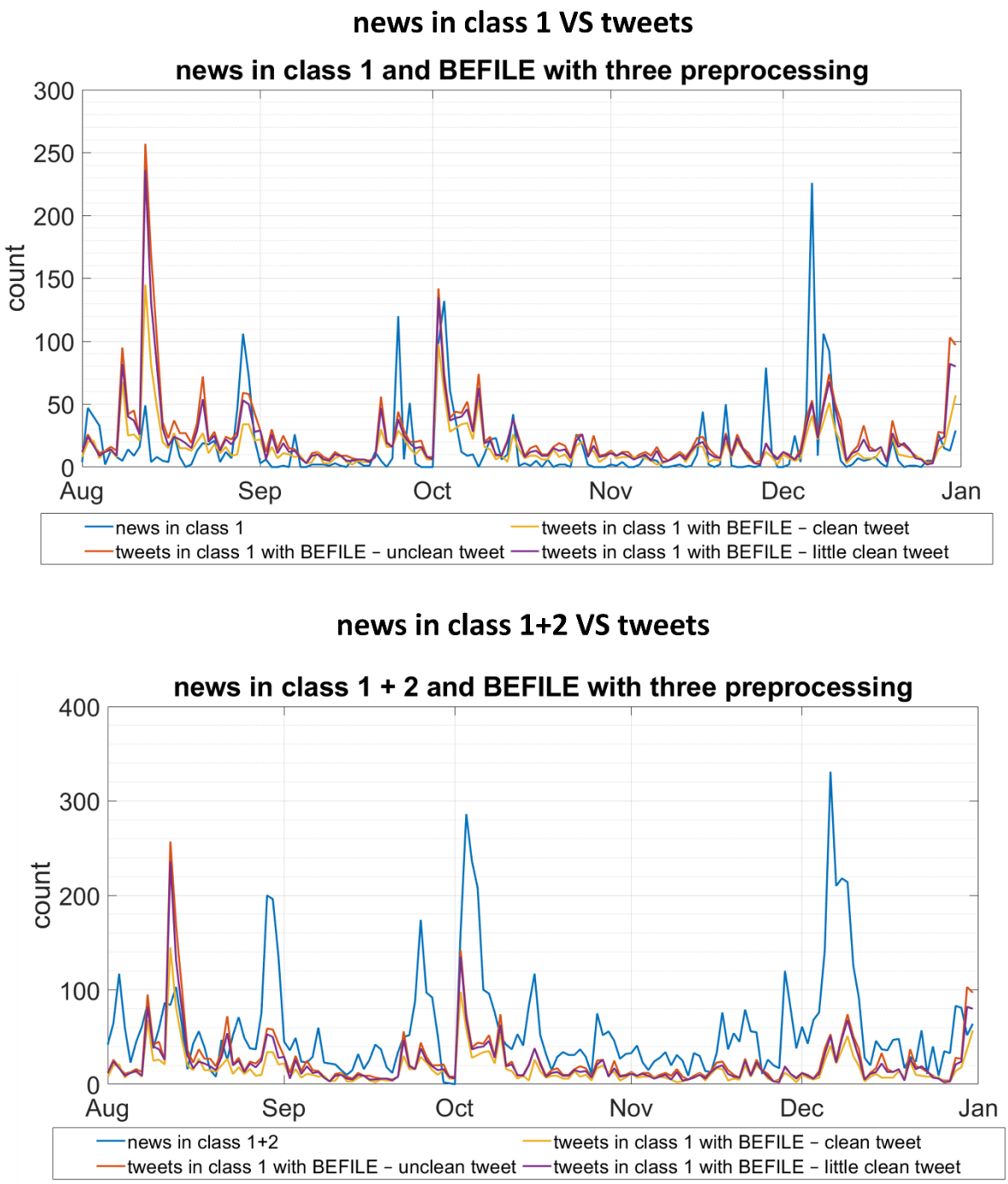

Figure 73: Daily distribution analysis between news in class 1 from Google News and tweets in class 1; then comparison between the distribution of news in classes 1 and 2 and tweets in class 1. Panels were generated using MATLAB R2021b.

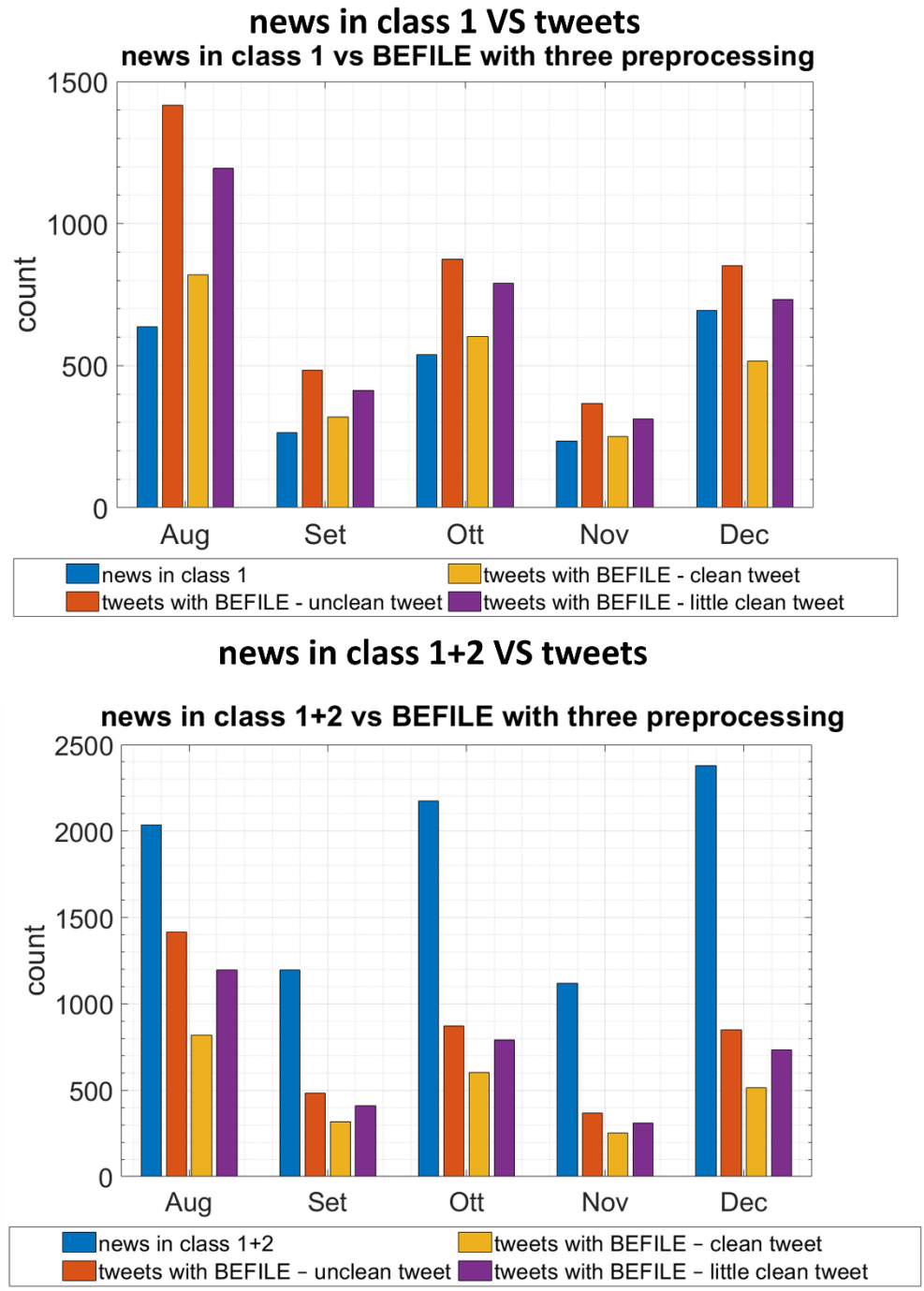

Figure 74: Monthly distribution analysis between news in class 1 from Google News and tweets in class 1; then comparison between the distribution of news in classes 1 and 2 and tweets in class 1. Panels were generated using MATLAB R2021b.

The correlation between targets in class 3 for news and class 0 for tweets has not been correlated because Twitter shows several noisy news items.

Subsequently, a more detailed analysis was conducted to obtain the correlation between the news in class 1 and in class 1+2 and classified tweets for each model of BEFILE (unclean text, little clean text). Figure 75 displays the correlation. Figure 75A, B and C indicate the news in class 1 and tweets in class

1 for each model. Figure 75D, E and F display the same correlation but regarding the news in class 1+2. Low values of count of news (“Landslide events” in the first three cases and “Landslide news” in the last three cases) and tweets corresponded. Sometimes, for low values of news and vice versa, it was not possible to outline a clear correlation.

The data do not follow a Gaussian distribution. For this reason, the R coefficient of Pearson and two nonparametric correlation indices were used to verify the rate of correlation between the analysed parameters. Kendall’s (K) and Spearman’s (S) rank correlation coefficients resulted in mean values, as reported in Figure 75 for each panel. High values have been calculated between news 1+2 and tweets. Considering the trend of the Pearson coefficient, Figure 75B and E show the highest point, in contrast with Figure 75A and D. The highest values of Kendal and Spearman are shown in Figure 75F.

Clearly, higher values have been reported for BEFILE with the preprocessing applied and considering the sum of classes 1 and 2 for the news.

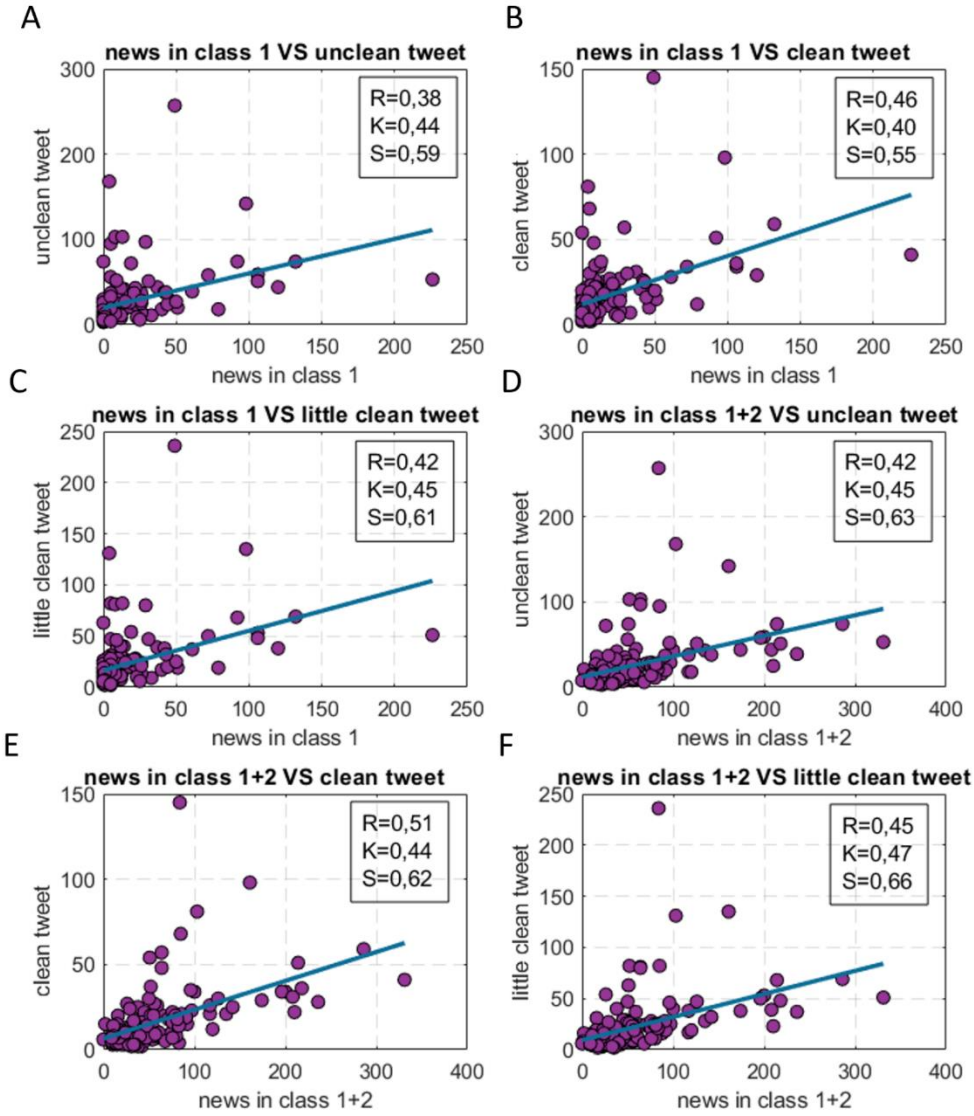

Figure 75: Detailed analysis between news in class 1 and classified tweets for each model. For each panel, the Pearson coefficient (R) and nonparametric values Kendal (K) and Spearman (S) were calculated. **A**, **B** and **C** show the correlation between news in class 1 and unclean tweets, clean tweets and little clean text. **D**, **E** and **F** display the correlation between the sum of classes 1+2 of news and different preprocessing applied to BEFILE. Panels were generated using MATLAB R2021b.

Due to privacy considerations, the geolocations of tweets are not available unless users actively elect to publish the information (Li et al., 2021). Twitter during extraction provides the entities from which the coordinates have been obtained. The model without preprocessing harvested more data with coordinates. To outline the spatial distribution, data from this model have been utilised. 18 data on 3993 include coordinates. Figure 76 shows the distribution. Each data point was examined to check the correct localization. In green, highlighted tweets describe landslide events with the right coordinates (in total are 7). In yellow, data are shown with approximate coordinates (in total are 3). For example, the coordinates of Basilicata refer to the region centroid. Another example is on Sicilia. The tweets on Sicilia published an event that occurred in Valtellina in Lombardia, but the coordinates

regarded the word “Ragusa” as localization. In red are tweets classified as useful about landslide events, but they include incorrect word associations (in total are 5).

Figure 76: Possible geolocalization of 18 tweets using entities. Tweets with coordinates not related to landslides are indicated in red. Tweets with coordinates not specifically of the event are represented in yellow. For example, one case in Sicilia showed a corrected classification, but the coordinates were localized on Ragusa (newsletter of the tweet). Entities that provided correct coordinates are depicted in green. The maps were generated using ESRI ArcGISPro.

As a second validation, the preview case study in Liguria was further analysed. Previous results consider data useful for creating a robust database for deep learning and are not exhaustive for a complete analysis. A new extraction was carried out to obtain a more exhaustive analysis. During data mining, retweets were regarded, expanding the classification dataset. A total of 6628 tweets were harvested on 24 November 2019. More than 4500 data points were extracted with respect to the previous database.

Figure 77 presents some tweet examples. To a certain extent, tweets provide good information details (such as in <<Breaking Italy News>>), while others publish along with text photo or video. Tweets from official channels have been collected, such as from firefighters, political branches, official newsletters (as <<TGr RAI Piemonte>>) and citizens. The last three types of tweets often present few or poor details of events and can be associated with answers or administrative discussion. However, these tweets are often after events and take on echo media impact functions.

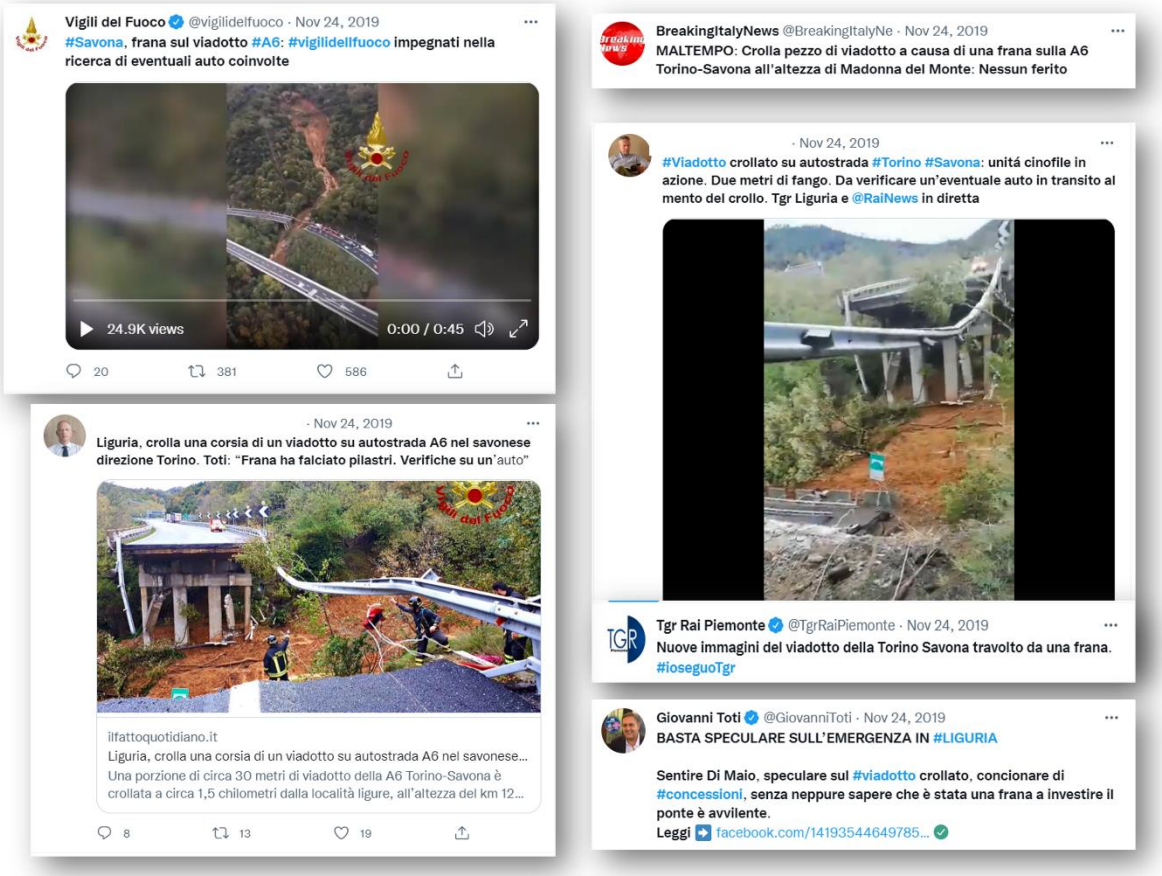

Figure 77: Some examples of Tweets about landslide events on the A6 motorway. Different types of details about events are possible to obtain, for example, place, time, feasible victims, damages, photo or video. Several types of users can post specific events: official channels (such as firefighters), citizens, politicians, and newsletters (such as TGr Rai Piemonte or BreakingItalyNews).

Based on the previous performance, the new dataset was subjected to little preprocessing before of classify in binary manner by BEFILE.

From the classified dataset, two panels have been compared, considering general tweets in the whole Italian territory (Figure 78A) and tweets with information about events (Figure 78B). Overall, in Italy on 24 November 2019, 2269 tweets described landslide events. Retweets were almost 138.397. Considering only landslide events in the A6 motorway, 1286 tweets were posted with retweets or a media impact of 114.663. The peak was measured at 14:00. The publishing tweets continued in the

next hours but with a decreasing trend. Analysing the Tweet datasets, it is possible to outline a sequence of events:

1. At 13:28, the first tweet that describes the fallen viaduct has been published.
2. At 13:36, the landslide-specific tweet (8 minutes later) has been released.
3. At 13:52, the tweet with a description of rescues in place was communicated (28 minutes after the event).

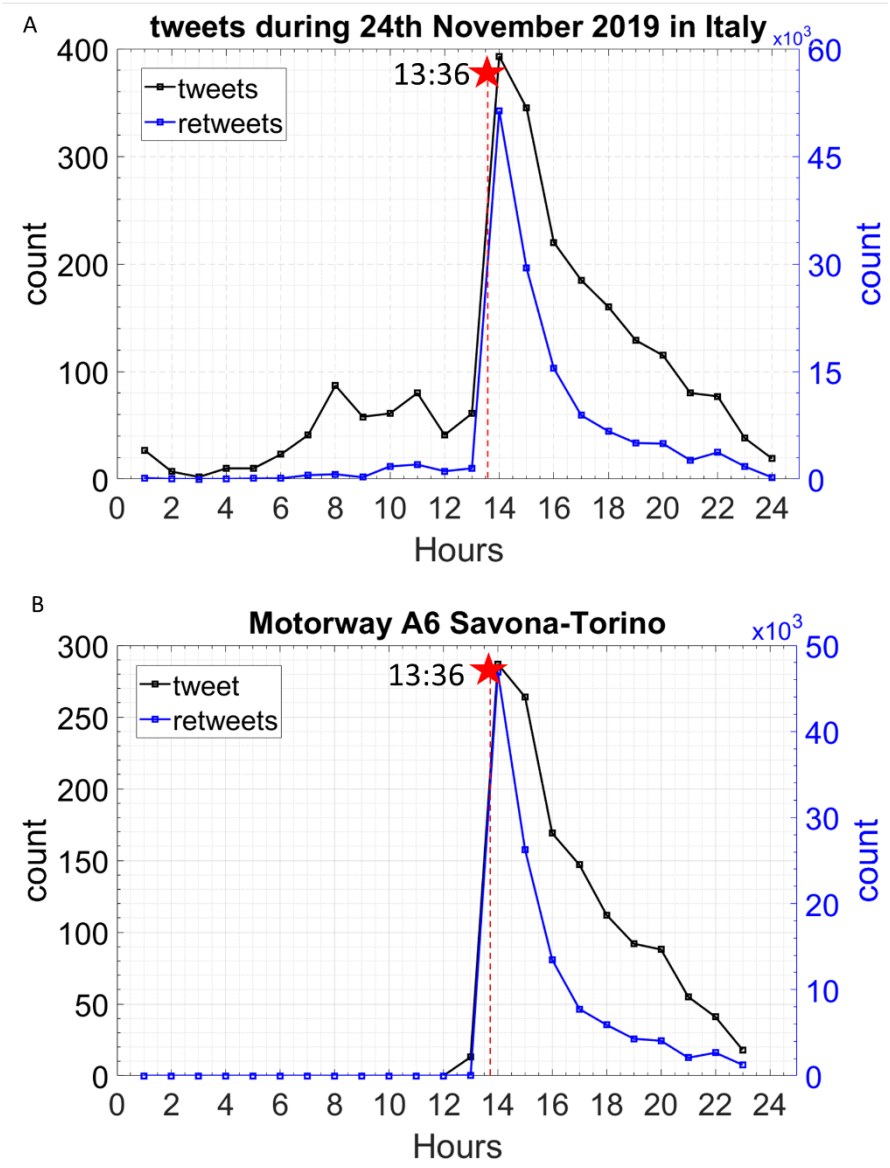

Figure 78: In **A** and **B,** data from Twitter are shown in two drafts: considering all tweets in Italy (**A**) and only tweets that describe the landslide event near motorway A6 (**B**). Both panels consider only 24 November 2019. Panels were generated using MATLAB R2021b.

The coordinates to the dataset classified by BEFILE were not signed manually. The entities have been utilized to obtain the geo-localization of the event. Unfortunately, only one data feature is shown by

latitude and longitude. Figure 79 shows the geolocated data in red. On the other hand, considering the ambiguity of information provided by entities, in this case, the coordinates present a good localisation of the event.

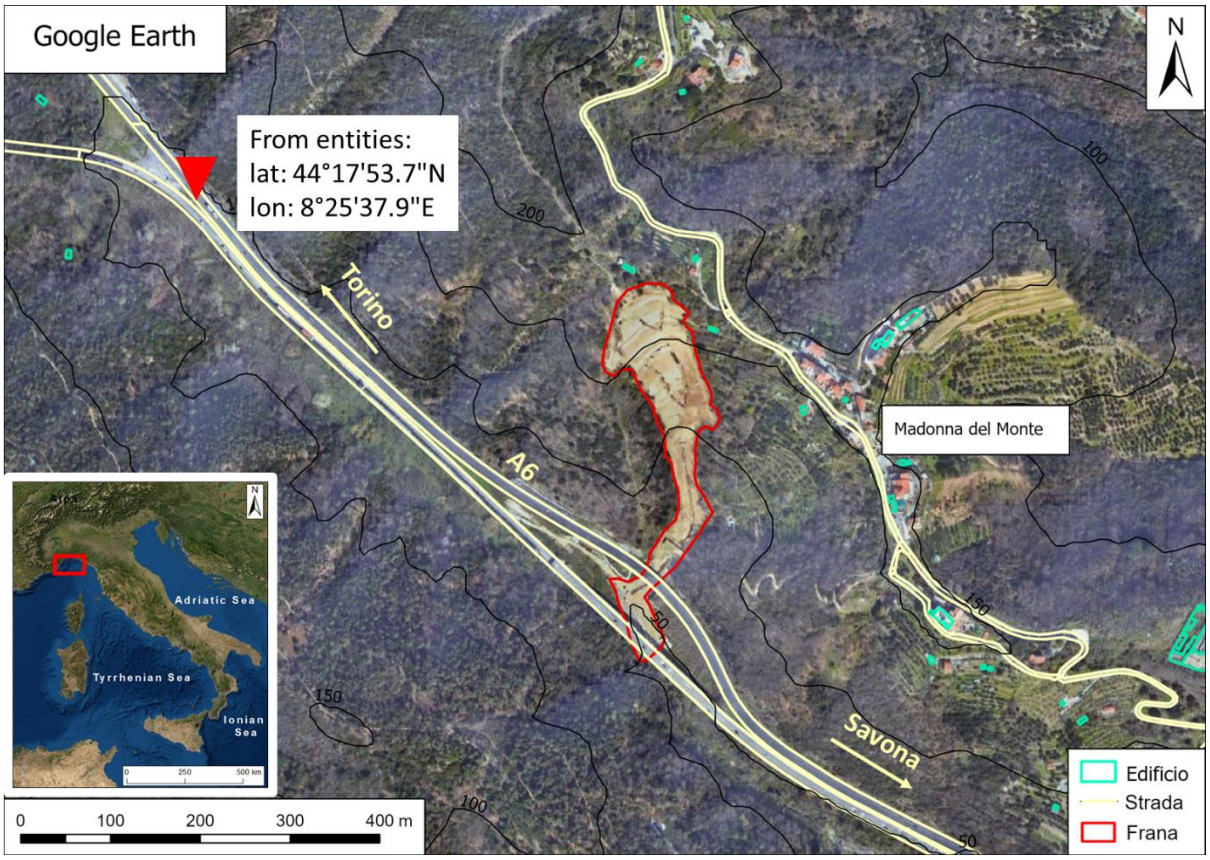

Figure 79: In the new dataset of 24th November 2019, one data provides the coordinates. The localisation of tweets using entities is highlighted in red. Considering the event and the point of the user or tweet text, the data present good event detection.

5 DISCUSSION

Different data sources were used to obtain information about landslide and flood events and their direct consequences in Italy over 10 years (2010-2019). The newspaper articles used were harvested by a data mining algorithm called the Semantic Engine to Classify and Geotagging News (SECaGN by Battistini et al., 2013). Data mining takes place within Google News, which considers national and local newspapers with more completeness (Franceschini et al., 2021). Over 10 years, in Italy, 184.322 articles related to landslides and 246.338 articles related to flood events were released by online newspapers. The retrieved articles were grouped based on the event they referred to. In this way, 32.525 landslide and 34.560 flood event news items were identified. Each news item was classified to create a landslide and flood database. The classification consists of two or three labels. Each label was based on relevance and spatial and temporal accuracy. This operation was necessary since each event can be reported from 1 or more newspapers based on its impact on the relevance of the affected area. For example, small landslides involving a major road or an important city can have a vast media echo. Conversely, landslides involving minor roads or small villages are reported only by local newspapers. Hence, the publication of one news item and its consequent mediatic impact in different newspapers depends on risk elements.

It is necessary to comment on the spatial resolution of the data used for landslide and flood events from Google News. Within newspapers, the exact location of an event is a parameter rarely available (Franceschini et al., 2021). Moreover, there is a claim that mass media attention is not uniformly distributed across disaster-affected areas (Fan et al., 2020). The presence/absence of news is affected by some factors, such as disruption in communication services, sociodemographic factors (the events affecting socially vulnerable populations receive less attention), the lack of exposed elements (roads, inhabited areas, etc., reporting low media impact) involved in landslides or the continued reactivation of one landslide over time (e.g., see La Saxe landslide; for each reactivation, more articles were published (Franceschini et al., 2022a)). The publication of one news item and its consequent mediatic impact in different newspapers depends on the event. Some areas can be involved with high-intensity precipitation that causes landslide phenomena with a low involvement of urbanized areas and therefore few published articles. On the other hand, a single rainfall event, even in an area with a low frequency of relevant rainfall events, can trigger a landslide with high human involvement.

Consequently, many articles can be published, creating a significant media impact. For these reasons, the use of newspaper articles may be useful for analyses over large areas but not to create a detailed landslide inventory or for detailed analyses (Franceschini et al., 2021).
For each dataset, a textual analysis was applied to obtain the word frequency. Interestingly, within "Landslide news" and "Flood news", similar and related words are present. For example, the words "frana" (landslide) and "allagamento" (flooding) are in both datasets, indicating how two events can be complementary or one subsequent to the other.

In the second step of this study, data mining was carried out within Twitter, and many analyses were performed to obtain an overview of tweets and their characteristics with respect to landslide events. The decision to extract information on landslide events was based on the lack of a state-of-the-art analysis of these natural hazards. A total of 13.350 data points were harvested, spread over 9 slots between 2011 and 2019. The dataset was subjected to a binary classification based on landslide information. Twitter offers many advantages regarding the speed of interaction between users but also some limitations about the spatial distribution and extractable tweet numbers.
In this work, the supervised deep learning method was used to obtain the data classified automatically. Several considerations were made to validate the tweet dataset but also to verify the feasibility of the classification model adopted.

5.1 Spatial and temporal distribution about landslide events

Landslide inventory from social media was used as a basis to obtain some information in terms of spatial and temporal distribution. Much information can be obtained based on the adopted classification. Within the landslide dataset, over 40% of the news items report useful information such as geolocalization and date (Class 1). Fifty-seven percent of the news items can identify the area involved in a landslide but not the date (Class 2). Both targets can be useful for analysing the distribution of landslide events and hence for estimating landslide hazards.
The regions and provinces with the most "Landslide news" (Class 1+2) are mainly located along the Alps and the Apennines. The geological, geomorphological context of the Alps, along with permafrost melting and frost–thaw cycles, leads every year to several landslides (Giardino et al., 2004; Ratto et al., 2007; Cignetti et al., 2016). Several areas in the Apennines also are highly involved in landslides. Considering only Class 1, the Liguria region demonstrates the highest number of events and daily landslides. Its territory is characterized by steep slopes with few flat areas along the coast and in the valley. The combination of significant urbanized areas and land use leads to a geomorphological

evolution characterized by an important presence of landslide and flood events. Flat areas are most likely to be affected by other geohazards, as floods are located along the northeast coast and in Puglia.

From a temporal point of view, the "Landslide events" increased from 2015 to 2019. The average number of days with landslides increased from 3 in the 2010-2014 period to 4 in the 2015-2019 period. The trend can be due to several causes. Some authors have related global climate change to the rise of global temperatures with a more frequent occurrence of extreme events (Rebetez et al., 1997; Easterling et al., 2000; Rosenzweig et al., 2008; Knight and Harrison, 2009; Keiler et al., 2010). In addition, inaccurate land use management can lead to an increase in mass movements throughout Italy over time (Italian National Institute for Environmental Protection and Research (ISPRA), 2020a). Landslide events had a certain seasonal distribution over the 10 years. In the wet season, the events were more frequent. Conversely, during the dry season, the frequency was lower. These results are in agreement with the literature, particularly for Campania (Cascini et al., 2014) and Toscana (Rosi et al., 2012) or nationally (Guzzetti et al., 2005; Calvello and Pecoraro 2018). In February and March, many landslides were reported. This trend can be associated with snowmelts, which occur as winter ends and temperatures rise. Snowmelts are a well-known landslide triggering factor in Italy (Cardinali et al., 2000).

The years 2013 and 2014 had an important number of days with landslide events. The reason is that over a long time interval, the La Saxe landslide suffered several reactivations. Mont de La Saxe, in Valle d'Aosta, had a rock fall-type landslide that caused damages or led to road closures. For each reactivation, new articles were published, and more days and high media impacts were recorded.

Subsequently, the landslide database was used as a proxy to correlate with other data sources. The attempt was to show how the combination of different data sources can be used to assist government authorities. Such embedding provides additional information for better knowledge of the landslide hazard of an area. Overall, four datasets with different information were explored: i) online newspapers, ii) rainfall data, iii) populations at risk for landslides and floods in Italy (Polaris database) and iv) earmarked funds for remediation work by the National Repository of Soil Defence interventions (ReNDiS database). In addition, the total number of published articles for each event was considered to outline the media impact or intensity of the landslide event. Finally, several ISPRA maps were used to validate the spatial distribution: percentage of hazard area, percentage of people at risk and percentage of buildings at risk.

The identification of factors controlling landslide distribution and occurrence is difficult because the relationship between landslides and their causative components varies spatially and temporally (Zhou et al., 2002). Nevertheless, a full understanding of these factors is relevant for the assessment of

natural hazards (Borgomeo et al., 2014) and their direct effects in terms of human lives and earmarked funds for soil protection. For this reason, rainfall data were analysed and correlated at different scales to “Landslide news” and “Newspaper articles”.

The regions most affected by landslides were mainly in northern Italy, along the Alps and in western regions, which is in agreement with their rainfall distribution and injure, death and missing (IDMs) numbers. Such distribution confirmed the studios by Neumayer et al., 2014 which supports that in countries of larger economic size will have more wealth potentially destroyable and are therefore expected to experience larger losses. On the other hand, the earmarked funds (by the central government) for soil protection outlined an inverse distribution. This appeared to be more widespread in southern Italy than in northern Italy. The fund distribution for soil protection depends on different variables: local, national and international political scenarios, social capital and investments by private actors such as citizens or environmental associations. The trend can be explained by the outcomes of the work of the World Bank and the United Nations (2010) and Padli et al. (2018). The authors hypothesized that regions with lower social capital (such as the southern regions of Italy) also may have weak economic structures. These can experience difficulties in securing adequate resources to recover from the damage of natural disasters. Campania, Sicilia, Puglia, Basilicata and Calabria exhibit a significantly lower index of economic well-being than the northern regions (Murias et al., 2012). The same regions, however, revealed prominent values of their percentages of infrastructure at landslide risk (Legambiente, 2021). Therefore, this may have led to a sharp increase in prevention activities for soil protection in recent years. For example, Campania exhibited the highest number of buildings at risk, in coherence with a high value of hazardous area. As expected, it was the area with the most funds allocated for soil protection. Another example, the Basilicata region, revealed low percentages of hazard areas, in contrast with its geological characteristics. In fact, it consists of land that is easily subject to erosion and runoff. Consequently, the loss of vegetation and land cover has led to serious instability phenomena. To address this issue, the region has opted for a policy of prevention and rehabilitation and an afforestation and hydraulic–forestry rehabilitation programme (De Stefano 2002). In both examples, the distribution of earmarked funds for soil protection is coherent with the goals of prevention and the recovery of damages caused by landslide events.

The central regions of Italy presented high values of “Landslide news”, “Newspaper articles” and frequency of relevant rainfall events. This aspect has been related to the Apennine chain, which crosses the country from north to south and is mainly formed by arenaceous flysch (Rosi et al., 2021; Vai et al., 2001; Rosi et al., 2018) in areas historically affected by landslides. In general, Liguria, Lombardia, Campania, Sicilia, Toscana and Emilia Romagna were the regions with the highest numbers of “Landslide news”. Puglia with 202 and Basilicata with 402 were the regions with the fewest publications. This trend was in agreement with the elevated values of hazardous areas as a function of

regional size, except for Liguria and Sicilia. Furthermore, such analysis confirmed that communities with higher resilience capacity, which are characterized by better social–environmental conditions, tend to have higher social media or crowdsourcing platforms use (Wang et al., 2021).

The divergent distribution of some variables in the Friuli Venezia Giulia, Trentino Alto Adige, Umbria, Puglia, Basilicata, and Calabria regions is linked to the occurrence of localized and very intense or sometimes extreme precipitation. Extreme weather events can trigger landslides in uninhabited areas, causing low media impacts and IDMs. Otherwise, as in the case of Umbria, Lombardia and Trentino Alto Adige, one single event or a few events can outline high IDMs. Some authors (Easterling et al., 2000; Rosenzweig et al., 2008; Knight et al., 2009; Keiler et al., 2010) have assumed that the increased occurrence of extreme events, even localized events, is caused by climate change. Loayza et al. (2012) recently stressed that natural disasters cause significant economic and physical losses, whose effects could spread beyond the immediate locality.

The combination of different data sources at a detailed scale can enhance the awareness of disaster managers for the aims of civil protection. There are 158 Warning Hydrogeological Zones (WHZs) that divide Italy on the basis of morphology, catchment boundaries and administrative limits. An analysis was applied to obtain more details about the spatial distribution of news and relevant rainfall events. A good correlation can be recognized between “Newspaper articles” and event counts but not with the frequency of relevant rainfall events. The absence of a correlation can be due to intrinsic characteristics in the news publications.

Each Italian region experienced some landslides in the investigated period, with approximately 1477 IDEMs per year. According to preview considerations, 2014 also was the year with important amount of rainfall, with 1007 mm/year and 3406 IDEMs spread among 19 regions involved in landslides. Based on the Polaris 2014 report, 2014 included several landslide phenomena that involved large areas. On 19-20 January, two different weather perturbations affected several zones of the Liguria and Emilia Romagna, causing death, injuries and damages in the railway network. On 3 May, an area in the Marche region was affected by intense rainfall, triggering many landslides and causing damage. The same scenario occurred on 2 August in some provinces in Veneto. From 3 to 6 September, in the northern part of Puglia, approximately 600 mm of rainfall was recorded, triggering several debris flows and mud flows. This amount of rain was very significant considering that the mean annual rainfall of these areas is approximately 800 mm/year. From 9 to 15 October, many provinces of Liguria, Toscana, Emilia Romagna, Piemonte and Friuli Venezia Giulia were affected by the same perturbations. Many landslides were triggered, causing damage and human losses. From 10 to 15 November, a similar meteorologic event occurred in northern Italy, involving the provinces of Liguria, Lombardia and Piemonte and causing more damage.

The highest number of “Landslide news” was recorded in 2019 (with an annual average of 904 mm/year). The mean annual rainfall was 12% higher than the average rainfall of the 1961-2019 climatic reference period (ISPRA, 2020b). In the same year, several events with long temporal distributions and involving large areas also were reported by the Polaris report (Polaris, 2019). Between 11 and 12 June 2019 in the Lombardia region, an extreme rainfall event (characterized by 125,6 mm in 12 hours) led to the triggering of many landslides. From 19 to 22 October, Lombardia, Liguria and Piemonte were involved in a very heavy intensity storm causing many landslides, including debris flows. Consequently, there was damage to infrastructure as well as victims and dozens of evacuees. In summary, 2019 can be referred to as the second year with the highest values of IDEMs and rainfall events, 2775 and 4529, respectively.

Generally, the temporal distribution of “Landslide news” revealed two increases, from 2010 to 2014 and then from 2015 until 2019. A similar trend was confirmed by Franceschini et al., (2022) who showed that the average number of days with landslides increased from 3 to 5 after 2014. Rainfall data followed a different distribution; in fact, rain data recorded a decrease from the first quinquennium to the second one. Conversely, the distribution and number of victims remained constant over the 10 years. These results are partially in accordance with the outcomes of Crozier (2020), the UN (2015) and Porfiriev (2016). The authors highlighted an increasing trend in the number of natural disasters and significant intensity rainfall events, with a consequent increase in the proportion of natural hazards, damages, and losses. These results agree with ReNDiS data, with which it is possible to derive the year of the intervention financings. It is reasonable to argue that funds for the events were distributed in years after the landslide. For example, the increase of earmarked funds from 2015 to 2017 can be referred to previous events (e.g., those happened in 2014, as in Campania, Emilia Romagna, Lazio, Liguria, Lombardia, Marche, Piemonte and Toscana).

Finally, a textual analysis was applied to obtain the frequency of words within headlines. Some words in Class 1 refer to synonyms of the “landslide”. In Class 2, the words refer to a hazard, alert, weather forecast or past or future event. In Class 3, the words are wrong associations or slang. Some words from Class 1 have been used. Given the ambiguity of some words (such as “maltempo”, “strada”, and “chiusa”), only some have been used as keywords, such as “frana” and “smottamento”.

5.2 Spatial and temporal distribution about flood events

In the flood dataset, over 14% of the news reported useful information (Class 1). Most of the news was classified as Class 2. The reason is linked to dilatation during the time and space of the event. This is in

contrast with landslide events, which are punctual. In any case, both datasets were used as inventories of landslide and flood events in the whole Italian territory during 2010-2019. The lack of Class 3 news is because incorrect word associations with “alluvione” or “allagamento” are less frequent than those with “frana” and its declinations.

The regions and provinces with more “Flood news” (Class 1+2) are mainly located along coasts, internal alluvial plains and along the main rivers. Considering only Class 1, the Sicilia region showed the highest number of events, consequences and days with floods. Its territory is characterized by high variability: mountains present a northern and hilly landscape in the north, south and west, and the eastern areas feature the widest plain and the large volcanic complex of Mount Etna (3346 m s.l.m) (Grauso et al., 2008) with poor natural vegetation. Most arable land is nonirrigated and located in hilly, sometimes steep areas where support practices are seldom applied. These characteristics make the Sicilian territory particularly vulnerable to erosion and soil degradation processes (Giordano et al., 2002).

As with the landslide database, the flood database was used as a proxy to correlate with other data sources. The same data sources were used but aimed to analyse flood events. Other correlations have not been carried out because the focus of this work is to obtain further information from landslide events.

The regions most impacted by floods were mainly in central and central-southern Italy. This trend agrees with the rainfall distribution, IDM numbers and earmarked funds for soil protection. In general, Valle d’Aosta, Piemonte, Liguria, Toscana, Emilia Romagna, Umbria, Marche, Abruzzo, Lazio, Molise, Basilicata and Trentino Alto Adige were the regions with good coherence between variables. The divergent distribution of some variables in Lombardia, Veneto, Campania, Sicilia, Puglia, Calabria and Sardegna is linked to the occurrence of localized and very intense or sometimes extreme precipitation. In some regions, the geomorphologic and geologic conditions also influence the event. In fact, Campania, Sicilia, Puglia, Calabria and Sardegna show similar conditions: hilly areas with poor vegetation. Lombardia and Veneto are regions characterized by wide plain areas in the south and important hydrographic networks in the north.

From a temporal point of view, the “Flood events” increased from 2016 to 2019. The trend can be due to several causes, as with increasing landslides. The “Flood events” had a certain seasonal distribution for the 10 years, as did the “landslide events” previously analysed. In October and November, many floods were reported. The trend can be associated with the autumn season and with a significant presence of rainfall.

The years 2014, 2017 and 2018 had an important number of days with flood events.

According to previous considerations, 2014 was also the year with important measures of rainfall, and 11 regions were involved in floods, causing almost 6697 IDEMs. The highest number of “Flood news”, along with a mean rainfall of 904 mm/year, was recorded in 2019.

In the same year, several events with long temporal distributions and involving large areas also were reported by the Polaris report. Between 11 and 12 June 2019 in the Lombardia region, an extreme rainfall event (characterized by 125,6 mm in 12 hours) led to the evacuation of over 1.100 people. From 19 to 22 October, Lombardia, Liguria and Piemonte were involved with very heavy intensity. In Piemonte, one victim, 3 injured and some evacuees were reported. From 11 to 19 November, central-southern Italy was affected by several rainfall events. During these events, 655 mm was measured at a rain gauge near Udine. Emilia Romagna and Toscana were the regions most involved. Over 3300 became evacuees. Venezia engaged in a catastrophic flood with a height of 187 cm, creating expensive damage to the infrastructure. From 22 to 25 November, intense precipitation fell between Liguria and Piemonte, with peaks of 500 mm/36 h and 420 mm/24 h, respectively. On 24 November, the landslide that involved motorway A6 was triggered. Three people were overwhelmed on a bridge by the Bormida River (Piemonte). After snowmelt, a flood event occurred on the Po River (Polaris, 2019). In summary, 2019 can be referred to as the second year with the highest values of rainfall events, 4529.

Generally, the temporal distribution of “Flood news” revealed two increases, from 2010 to 2014 and then from 2015 until 2019. A similar trend also has been evaluated for “Flood news”, and it has been reported for IDEMs. In contrast, rainfall and ReNDiS data followed a different distribution that recorded a decrease from the first quinquennium to the second one. Rain data can be described with an increasing trend of significant intensity rainfall events. The consequences included those reported by Crozier (2020), the UN (2015) and Porfiriev (2016) and the increased propension of natural hazards, damages, and human losses. Moreover, the spatial distribution issues observed for landslide events are the same for flood events.

Finally, a textual analysis was applied to obtain the frequency of words within headlines. Some words in Classes 1 and 2 refer to synonyms of the “flood”. Given the repeatability of some words, it is clear that flood events are spread over time and cannot be considered as punctual as landslides.

5.3 Data mining and BERT for landslide events on Twitter

The data mining technique allows us to obtain and create datasets for specific events. In this work, using Twitter as a data source and Python as a tool, it was able to obtain the tweet text referring to landslide events. Twitter is an excellent resource for event detection. People share opinions and information about the situation. Detecting situational tweets is a challenging task. The first step is to

collect some keywords to apply the data mining technique within Twitter. The keywords and period time of extraction were chosen based on the analysis of newspaper articles. However, even with the overwhelming amount of data that can be found on Twitter, often it is not enough to use keywords alone to obtain useful tweets (Nguyen et al., 2016).

The Twitter dataset comprises various slots with different temporal distributions. The main purpose is not to recreate the same inventory of landslide phenomena as for newspapers but rather to apply classification techniques. Therefore, the dataset is considered neither complete nor exhaustive for the 2011-2019 period. As demonstrated by Zhou et al. (2022), several issues may arise due to the nature of big social media data. Tweet analysis tends to favour those who use social media more often. Uneven usage of social media may lead to biased consequences. Moreover, social media posts suffer from locational bias, temporal bias, and reliability issues. Such issues should be considered while further analysing the spatiotemporal patterns of the identified tweets for detecting vulnerable communities or assessing disaster damages (Zhou et al., 2022). In this way, tweets can identify the most virtuous or resilient region with respect to the hazard event. In fact, user interaction can identify active social behaviour, inclined to information (transmitting and/or receiving it) and consequently resilient to the event.

The dataset was subjected to a binary classification based on landslide information. Disaster tweet classification studies can be considered natural language processing (NLP) tasks. Furthermore, 4158 data points were assigned approximate text-based coordinates. As with the previous results from news, the point distribution follows the main chain mountains (Alps and Apennines), although the data are not complete. The regions most involved in landslides are Liguria, Piemonte, Calabria, Veneto, Friuli Venezia Giulia and Trentino Alto Adige. All these regions, except for Friuli Venezia Giulia, also presented important values for “landslide new”. Liguria, Piemonte, Calabria and Trentino Alto Adige are characterized mainly by chain mountains and hills. Conversely, Veneto and Friuli Venezia Giulia feature chain mountains in the northern area and large areas of plains in the southern area. Such result is in agreement with studies by Wang et al., (2021), which identified communities with higher resilience capacity, which are characterized by better social–environmental conditions (see data from RENDiS), tend to have higher Twitter use.

In the second step, multiple statistical analyses and natural language processing were performed, leading to multiple considerations. The high propensity of tweets in Class 0 demonstrates the difficult handling and ambiguity that characterize the data from Twitter. Therefore, a strong filtering system must be applied to handle these data. From the natural language processing techniques applied, it was possible to delineate the occurrence of words and the distribution of the text of tweets by applying preprocessing. By applying data cleaning, it is possible to distinguish the preponderance of meaningless words in tweets classified in Class 0 from those in Class 1. By applying preprocessing and

considering double occurrences, it is possible to see that the word associations are more consistent. In fact, tweets classified as more informative note 'frana minaccia','strada chiusa' and 'traffico frana' descriptions. These refer to infrastructure and roads, showing that there is indeed more public attention. In addition, there is a preponderance of 'landslide traffic' and 'landslide weather'. More ambivalence should be attributed to the word 'landslide', which is very present in Class 0. This word is often associated with two distinct aspects: i) generic collapse of infrastructure without specification of the cause, which therefore is not actually classifiable as a landslide event; and ii) emotional, sentimental or situational terms, such as ‘sono una frana in matematica' or la ‘borsa frana' or ‘frana l'inter contro la fiorentina'. The same problems were encountered with the synonym 'landslip', which should contain more specifics, especially in the first case. This duplicity may be related to the inherent characteristic of the tweet text: the speed with which it is published and summarized, which shows a lack of information. This can be considered an advantage and a disadvantage. It can be an advantage in the sense that the messages are not long and it is easy to see the information they contain, but it can be a disadvantage if the user does not express himself or herself accurately in the text (Dragović et al., 2019). In this case, the text will be meaningless (Goswami et al., 2018).

The Tweet dataset was compared with the distribution of newspaper articles by Google News. Tweets were not classified on a temporal basis; therefore, Classes 1 and 2 of news were compared with tweets in Class 1. The trend showed a good temporal correlation. If the tweets presented a decrease almost immediately at the event, the news presented echoes in the next days. The reason can be linked to several aspects:

1. newspaper news needs more steps for publication than a tweet;
2. the event(s) present an impact distribution over many days, hence articles also are published in the following days;
3. the consequences of the event or the damage caused are felt in the following days, so the articles are published repeatedly over time.

The same procedure was applied to the noninformation data in both databases (Class 3 for news and Class 0 for tweets). The few news items in Class 3 verify the effectiveness of the filtering systems adopted in the design of SECaGN for newspaper articles and confirm the efficiency and utility of data cleaning. However, the tweet trend is clearly the opposite, showing noisy, filthy and useless data.

When analysing Twitter data for community resilience, tweet content, networks, or metadata are used. These data analysis techniques form the basis of social-sensing network (Fan et al., 2020; Kryvasheyeu et al., 2016, 2015), in which individuals are used as sensors which contribute to the knowledge gained about crisis event (Rachunok et al., 2021). A community or system's ability to sense

is a critical part of its resilience when coupled with anticipation, learning, and adaptation (Park et al., 2013). As a result of social media sensing's flexibility, analysis can be aligned with a community's ability to learn, anticipate, and adapt. By this point, the interaction by users can allow to obtain the parameter of resilience in agreement with the study of Dufty et al., (2012) and Wang et al., (2021). In fact, communities with higher resilience capacity, which are characterized by better social–environmental conditions, tend to have higher Twitter use.

To outline the efficiency of Twitter, some case studies were considered within the classified database. Within the tweet dataset, there was an event of a viaduct collapse due to a landslide in Liguria. The event had an important media echo because it was associated with the collapse of the Morandi Bridge, the lack of prevention and the failure of land recovery. There were no victims during the event. A number of available data sources were correlated: rainfall, news and tweets. The rainfall distribution was based on the analysis of data from the days before the event from the nearest pluviometer. The period was chosen based on the fact that the landslide triggering was caused by intense and short rainfall. The nearest pluviometer was located north of Savona, and measurements from 1 October 2019 to 24 November 2019 were considered, showing a considerable accumulation of rainfall. This choice is because other neighbours' pluviometers recorded nothing and most likely were not in operation. Such distribution was clear through the IDW.

The first tweet was recorded at the moment of the collapse at 13:36, and it was followed by many others with varying specifications. These data were compared with the publication of news items. The first article was published one hour after the event. Such comparison points to the remarkable speed of publication and dissemination in the crowdsourcing platform and the actual 'delay' in publishing an article. From tweet counts, it is possible to obtain some maps with a possible alert system. Two types of maps were analysed using coordinates from text and coordinates of the event. Better results in terms of distribution were obtained in the first case. The highest values were localized in the central region between Savona and Genova. This trend showed that an event does not have a point effect but also repercussions in the areas closest to it. This demonstrates how data on a municipal scale is in any case exhaustive in the civil protection phases.

The manually classified dataset provided a solid base for applying deep learning using the natural language technique. "**Bert For Information on Landslide Events**", or BEFILE, was created using transformer architecture. This script allowed us to classify text into two classes (0 and 1) based on landslide information in the Italian language. This analysis led to a considerable advancement of the BERT classifier, which until now was very often used for a variety of analyses in the English language and different fields. Two advantages resulted from this project: i) the Italian-language classified dataset for landslide events fills that present gap of analysing natural events using Twitter, which has not yet been exploited to a great extent for landslide events; and ii) BERT was trained to detect this

information and proved to be an excellent classifier for the Italian language for landslide events. Although such an aspect involves an issue, people using languages other than Italian on social media cannot be leveraged. One way to solve the problem is to train a corresponding model for the target language or scrutinize whether a unified model can render a reliable performance across numerous languages (Zhou et al., 2022).

Three tests of BEFILE were carried out to obtain text classification, changing the setup of the preprocessing. This procedure was necessary to outline the best setting of the data cleaning parameters. For each iteration, EarlyStopping was set to 15 to obtain the best performance on validation and to avoid useless iterations. Based on the results obtained for each model, the best trade-off is represented by the BEFILE without preprocessing. In fact, the first BEFILE showed important values of accuracy equal to 96% and an AUC of 0,95. To validate BEFILE, two case studies have been considered. The first study considers a comparison between newspaper articles and Twitter datasets for a part of 2020. The second study regarded a detailed analysis of the motorway falling in the Liguria region. In both cases, a new database from Twitter was extracted and classified using the model. Before applying BEFILE, different sets of preprocessing were applied to the new database.

For the first study, 39.780 data points were classified by BEFILE. To define the best detected data, a validation to a new Twitter dataset was applied using newspaper articles. A good correlation was found between news in Class 1+2 and tweets, also through nonparametric values. In addition, this study considered landslide information from the perspective of spatial analysis. Entities were leveraged to estimate the localization of events. The geolocations of tweets are not available unless users actively elect to publish the information. Although the information of a user's registration location can be used as an alternative, it might not directly connect to the location of an event observation because the location where a user posted a tweet can be different from the user's registration location (Li et al., 2021). In general, the spatial result reveals that only a small amount of data (18 of 3993) tweets presented coordinates associated with landslide events. Such a distribution may not be sufficient to support a reliable recovery assessment. For this reason, in the spatial distribution analysis, tweet data were not correlated with news.

The second case study, of the viaduct close to the Madonna del Monte in the Liguria region, was utilized as a second validation. Notably, for this event, two existing datasets presented different data distributions. Some limitations can be pointed out. In fact, in the first case, within the query of a request for the Twitter API, retweets were removed, entering such parameters in the script ("*-is:retweet"*). Conversely, in the second case, retweets were considered, removing such requests within the query. Moreover, the number of tweets with specific coordinates was small (1 for both datasets).

In general, all tweet datasets are not exhaustive for a further reason. Twitter provides a rate limit of extraction. Every day, many thousands of developers make requests to the Twitter API. To help manage

the sheer volume of these requests, limits are placed on the number of requests that can be made. The maximum number of requests that are allowed is based on a time interval, some specified period or a window of time. The most common request limit interval is fifteen minutes. If an endpoint has a rate limit of 900 requests/15 minutes, then up to 900 requests over any 15-minute interval are allowed. In this project, Twitter API v2 and OAuth 2.0 Bearer Token granted 300 requests for each 15-minute interval. This aspect reduces and limits the analyses and does not allow us to obtain a complete dataset for an event.

Considering the preview results, considerations can be made regarding preprocessing and the number of classified data. The nonapplication of preprocessing before the model resulted in a high classification of the data but also risked obtaining false data. On the other hand, the application of any preprocessing completely undermines the text, risking changing the context and meaning. This results in a considerable loss of data. Based on the performance of BEFILE, the model with middle preprocessing was chosen to classify the new dataset in a binary manner. This also was found to be a good compromise based on studies by Dharma et al. (2022). The author demonstrated that the use of stop word removal as an example can decrease the overall performance of the model. This is because stop word removal reduces the size of the dataset, and for a text, this often can change the overall meaning within the text, even though it reduces the training time. Therefore, the use of stop word removal is often not necessary, and having a larger dataset size is better for the model, as it can improve the overall performance of the model. Nevertheless, BEFILE is located between the works by Madichetty et al. (2020) and Dharma et al. (2022), which use coupling techniques between CNN and BERT embedding. However, from a practical point of view, this study provides useful perspectives for decision-makers to consider when using social media as an additional information resource for rapid damage assessment. BEFILE makes possible the detection of landslide events within tweets and brings state-of-the-art integration in NLP technology of text classification. At the same time, several problems may arise due to the nature of big social media data analysis and some limitations of this research. These problems should not be ignored when translating the research results into practice.

6 CONCLUSION

Mass media is generally the first and primary source of information about hazards for the public (Fischer, 1994). The use of data mining techniques is advancing in different ways. The main aim of this work is to demonstrate the utility and capability of social media to detect events in areas without physical sensors that would detect natural hazard directly. Different steps of analyses have been applied to define spatial and temporal distribution of newspaper articles for whole Italian territory. Such analysis have allowed to outline a form of resilience in function of number of articles published during and after the event. Below many steps are described for demonstrating as it is possible manage the data and what it is possible to derive. In this study, different data sources were analysed to obtain information about natural events (landslides and floods) at the national scale. In the first step, news of landslides and floods was analysed using as source the Multi-risk Information Gateway or MIG platform, which collected articles about natural events (landslides and floods) at the national scale from Google News. For both kinds of events, 10 years were analysed. In total, 32.525 landslide news items and 34.560 flood news items were collected. The datasets are classified into classes based on the thematic, temporal and spatial relevance of the news. This classification makes it possible to outline the temporal and spatial distribution of the events, their media impact and also to outline a hazard map on a regional and provincial scale. Different aspects of newspaper distribution can be obtained: “Landslide and Flood news”, which outlines the hazard areas, and “newspaper articles”, which describe the media impact or event impact. The integration of natural hazard information and social media could improve warning systems to enhance the awareness of disaster managers and citizens about emergency events. To reduce the gap between social media and traditional sensors several correlations were applied. News was correlated to rainfall data and event effects in terms of victims (POLARIS) and earmarked funds (REnDIS).

The spatial distribution revealed that there are more “Landslide news”, “Newspaper articles”, and IDEMs in the northern regions and in some cases in southern regions (Campania and Calabria). A similar trend was found in the frequency of relevant rainfall events. Conversely, the distribution of earmarked funds is more concentrated in southern Italy than in northern Italy. The increase in prevention activities for soil protection in recent years, in southern Italy and partially in central Italy, can be linked to the high percentages of landslide hazards and buildings at risk that characterize them.

“Landslide news” showed an increasing trend from 2010 to 2014, and it repeated in the 2015-2019 period, in contrast with rainfall data, “Newspaper articles” and reported expenses, while the IDEMs number remained constant.

The regions most impacted by floods were mainly central and central-southern Italy. This trend agrees with the rainfall distribution, IDM numbers and earmarked funds for soil protection. The Sicilia region showed the highest number of events or “Flood events” and days of consequence with floods. From a temporal point of view, the “flood news” increased from 2016 to 2019. The trend can be due to several causes similar to those with increasing landslides. Other correlations have not been carried out because the focus of this work is to obtain further information on landslide events. Given the present literature on data mining for flood events and the absence of studies on landslide events, the analysis focused on the latter events to deepen and analyse a topic not truly addressed in social media analyses and crowdsourcing platforms.

In the second step of this work, a new data mining technique in Twitter was applied using appropriate keywords extracted by newspaper headlines. Several techniques have been developed for data mining in social media for many natural events, but they have rarely been applied to the automatic extraction of landslide events. This makes it possible to fill the gap in the literature with respect to landslide events. One script was set to obtain the database from Twitter. The data mining technique has thus far been applied to newspaper news, but now, with the appropriate use of keywords, also within the Twitter dashboard. Twitter is an excellent resource for event detection. The dataset, from Twitter, features by 13.349 data, was classified manually, providing a solid base for applying deep learning.

A wide range of natural language processing use cases exist, and disaster tweet classification can be considered one of them. Exploring the dataset, some case studies have been analysed. Based on tweet counts, possible alert system maps were created. These results demonstrate how the data on a municipal scale is in any case exhaustive in the civil protection phases.

The classification allowed us to identify the most relevant tweets in terms of the temporal and spatial accuracy of landslide event identification. The harvested dataset was classified manually, providing a solid base for applying deep learning. Moreover, the Italian-language classified dataset for landslide events fills that present gap of analysing natural events using Twitter. This method has not yet been exploited to a great extent for landslide events. A script was created for text classification using the transformer architecture with the BERT method. “**Bert For Information on Landslide Events**”, or BEFILE, allows the classification of text into two classes (0 and 1) based on landslide information in the Italian language. This analysis leads to a considerable advancement of the BERT classifier, which until now was very often used to analyse data in the English language for different fields. BEFILE without preprocessing showed significant accuracy, equal to 96% and an AUC of 0,95, locating itself between implementing models with CNN. BEFILE showed promising results in classifying and thus detecting

information about landslide events, but some limitations must be considered. Datasets can be considered not complete mainly for some respects: i) non-exhaustive keywords, ii) Twitter limits the number of extractions per time unit and ii) the lack of geolocation of data.

Despite the limitations of social media data with respect to validated official reports, this study confirms that relevant and statistically significant information on landslide and flood hazards can be obtained by data mining of social networks during emergencies. Such data, properly filtered and classified, may be of notable help in increasing our present capability of calibrating and validating early warning models, with particular reference to data-scarce areas and back-analysis of undocumented past events. The information collected from social network whilst an adversity or crisis event recount a bottom-up symptom of the effects of a tragedy or crisis as it's felt by the human beings in a community. Social network information can bolster community suppleness analyses by functions as an information source which is closely aligned with the spatial and temporal scales calamity and crisis decision making. Some evaluations can represent a useful tool to understand and assess the impact of natural disasters, as well as to plan the best strategies for risk reduction at regional or national scale. Furthermore, it was demonstrated as Twitter can be utilized as source of rapid information and detection for landslide event. A possible contribution about implementation of specific communication and warning guidelines with respect to natural events such as landslides has been proposed. Creating a simple homogeneous language can available the communication between decision-makers and citizens, but also decision-makers and data analysis-makers. From a practical perspective, this study provides useful perspectives for decision-makers to consider when using social media as an additional information resource for rapid damage assessment.

BIBLIOGRAPHY

Aburn, G., Gott, M., & Hoare, K. (2016). What is resilience? An integrative review of the empirical literature. Journal of advanced nursing, 72(5), 980-1000.

Acar, A., & Muraki, Y. (2011). Twitter for crisis communication: lessons learned from Japan's tsunami disaster. International Journal of Web Based Communities (IJWBC), Vol. 7, No.3, 2011. https://doi.org/10.1504/IJWBC.2011.041206

Agliardi, F., & Crosta, G.B. (2003). High resolution three-dimensional numerical modelling of rockfalls. International Journal of rock mechanics and mining sciences. Volume 40, Issue 4, Pages 455-471. https://doi.org/10.1016/S1365-1609(03)00021-2

Agostini, A., Tofani, V., Nolesini, T., Gigli, G., Tanteri, L., Rosi, A., Cardellini, S., & Casagli, N. (2014). A new appraisal of the Ancona landslide based on geotechnical investigations and stability modelling. Q J Eng Geol Hydrogeol 47:29–43.

Alam, F., Qazi, U., Imran, M., & Ofli, F. (2021). HumAID: Human-Annotated Disaster Incidents Data from Twitter, In ICWSM, 2021.

Alam, F., Imran, M., & Ofli, F. (2019). "CrisisDPS: Crisis Data Processing Services," in In Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM), 2019, Valencia, Spain, 2019. [Online]. Available: http: //www.wis.ewi.tudelft.nl/twitcident/

Alam, F., Ofli, F., Imran, M., & Aupetit, M. (2018). A twitter tale of three hurricanes: Harvey, Irma, and Maria. In Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Rochester, NY, USA, May 20-23, 2018.

Alaparthi, S., & Mishra, M. (2021). BERT: A sentiment analysis odyssey. Journal of Marketing Analytics, 9(2), 118-126.

Alharbi, A., & Lee, M. (2019). Crisis Detection from Arabic Tweets. In Workshop on Arabic Corpus Ling., 72–79.

Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

Avvisati, G., Sessa, E.B., Bellucci, E., Colucci, O., Marfè, B., Marotta, E., Nave, R., Peluso, R., Ricci, T., & Tomasone, M. (2019). Perception of risk for natural hazards in Campania Region (Southern Italy). Int. J. Dis. Risk Red. 2019, 40, 101164.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In: Proceedings of the International Conference on Learning Representations. San Diego.

Barriere, V., & Balahaur, A. (2020). Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 266–271, Barcelona, Spain (Online).

Barbieri, F., Anke, L.E., & Camacho-Collados, J. (2021). XLM-T: A multilingual language model toolkit for twitter. arXiv preprint arXiv:2104.12250.

Baranowski, D.B., Flatau, M.K., Flatau, P.J., Karnawati, D., Barabasz, K., Labuz, M., Latos, B., MS Jerome, M.S., Paski, J.A., & Marzuki (2020). Social-media and newspaper reports reveal large-scale meteorological drivers of floods on Sumatra. Nature communications, 11(1), 1-10.

Basile, V., Barbieri, F., Croce, D., Nissim, M., Novielli, N., & Patti, V. (2016). Evalita 2016 Sentipolc Task - Task Guidelines. pages 1–12.

Battistini, A., Segoni, S., Manzo, G., Catani, F., & Casagli, N. (2013). Web data mining for automatic inventory of geohazards at national scale. Applied Geography 147-158.

Battistini, A., Rosi, A., Segoni, S., Lagomarsino, D., Catani, F., & Casagli, N. (2017). Validation of landslide hazard models using a semantic engine on online news. Applied Geography 59-65.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Hong Kong. 3613– 3618

Bengio, Y., Ducharme, R., Vincent, P., et al. (2003). A neural probabilistic language model. J Mach Learn Res, 3: 1137–1155

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell, 35: 1798–1828

Bernier, S. (2013). Social Media and Disasters: Best Practices and Lessons Learned. Disaster Preparedness Summit, American Red Cross.

Bianchini, S., Raspini, F., Solari, L., Del Soldato, M., Ciampalini, A., Rosi, A., & Casagli, N. (2018) From picture to movie: twenty years of ground deformation recording over Tuscany region (Italy) with satellite InSAR. Front EarthSci 6:177.

Birant, D., & Kut, A. (2006). ST-DBSCAN: An algorithm for clustering spatial-temporal data. Science Direct. Data & Knowledge Engineering 60 (2007) 208-221 doi:10.1016/j.datak.2006.01.013

Birkbak, A. (2012) Crystallizations in the Blizzard: Contrasting informal emergency collaboration in Facebook groups. In: Malmborg L., Pederson T. (Eds.), Making Sense Through Design. Proceedings of the 7th Nordic Conference on Human-Computer Interaction, Copenhagen, Denmark, 428–43710.1145/2399016.2399082

Blanford, I.J., Bernhardt, J., Savelyev, A., Parodi, G.W., Carleton, A.M., Titley, D.W., & MacEachren, A.M. (2014). Tweeting and Tornadoes. In: Hiltz S.R., Pfaff M.S., Plotnick L., Shih P.S. (Eds.), Proceedings of the 11th International ISCRAM Conference. University Park, Pennsylvania, USA, 319-323

Borgomeo, E., Hebditch, K.V., Whittaker, A.C., & Lonergan, L. (2014). Characterising the spatial distribution, frequency and geomorphic controls on landslide occurrence, Molise, Italy. Geomorphology, 226, 148-161.

Brunkard, J., Namulanda, G., & Ratard, R. (2008). Hurricane katrina deaths, louisiana, 2005. Disaster medicine and public health preparedness, 2(4), 215-223.

Burel, G., Saif, H., Fernandez, M., & Alani, H. (2017). On semantics and deep learning for event detection in crisis situations. In Workshop on semantic deep learning (SEMDEEP), ESWC 2017.

Burton, M.L., & Michael, J.H. (2005). Hurricane Katrina: Preliminary Estimates of Commercial and Public Sector Damages. Unp. Manuscript. Center for Business and Economic Research, Marshall University, Huntington.

Buscaldi, D., & Hernandez-Farias, I. (2015). Sentiment analysis on microblogs for natural disasters management: a study on the 2014 genoa floodings. In Proceedings of the 24th international conference on world wide web (pp. 1185-1188).

Calvello, M., & Pecoraro, G. (2018). FraneItalia: a catalog of recent Italian landslides. Geoenviron Disasters 5(1):1–16

Campobasso, C., Delmonaco, G., Dessì, B., Gallozz, P.L., Porfidia, B., Spizzichino, D., Traversa, F., & Vizzini, G. (2013). Long term strategies and policies for geological and hydraulic risk mitigation in Italy: The ReNDiS project. In Landslide science and practice pp. 39–45. Springer, Berlin, Heidelberg (2013).

Cardinali, M., Ardizzone, F., Galli, M., Guzzetti, F., & Reichenbach, P. (2000). Landslides triggered by rapid snow melting: the December 1996-January1997 event in Central Italy. Mediterranean Storms (Proceedings of the EGS Plinius Conference held at Maratea, October 1999), Italy

Carley, K.M., Malik, M., Landwehr, P.M., Pfeffer, J., & Kowalchuck, M. (2016). Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia. Saf. Sci., 90, 48–61.

Cascini, L., Sorbino, G., Cuomo, S., & Ferlisi, S. (2014). Seasonal effects of rainfall on the shallow pyroclastic deposits of the Campania region (southern Italy). Landslides 11(5):779–792

Castillo, C. (2016). Big crisis data: Social media in disasters and time-critical situations. Cambridge University Press.

Cavallo, E., & Noy, I. (2009). The Economics of Natural Disasters: A Survey, IDB Working Paper Series, No. IDB-WP-124, Inter-American Development Bank (IDB), Washington, DC

Chatfield, A.T., & Brajawidagda, U. (2013). Twitter Early Tsunami Warning System: A Case Study in Indonesia's Natural Disaster Management. In: Proceedings of the 46th Hawaii International Conference on System Sciences, IEEE Computer Society, Maui, Hawaii, 2013, 2050 – 2060, 10.1109/HICSS.2013.579

Chaturvedi, A., Simha, A., & Wang, Z. (2015). ICT infrastructure and social media tools usage in disaster/crisis management. Proceedings of the Regional Conference of the International Telecommunications Society (ITS), Los Angeles, 2015, 1-33Search in Google Scholar

Cignetti, M., Manconi, A., Manunta, M., Giordan, D., De Luca, C., Allasia, P. & Ardizzone, F. (2016). Taking advantage of the ESA G-POD service to study ground deformation processes in high mountain areas: a Valle d'Aosta case study, Northern Italy. Remote Sens 8(10):852

CNN: Louisiana's mammoth flooding: By the numbers, 2016 https://edition.cnn.com/2016/08/16/us/louisiana-flooding-by-the-numbers/index.html

Cobo, A., Parra, D., & Navón, J. (2015). Identifying relevant messages in a twitter-based citizen channel for natural disaster situations. Proceedings of the 24th international conference on world wide webWWW'151189–1194.

Colle, B.A., Westrick, K.J., & Mass, C.F. (1999). Evaluation of MM5 and Eta-10 precipitation forecasts over the Pacific Northwest during the cool season. Weather Forecast. 14(2), 137–154

Collobert, R., Weston, J., Bottou, L., et al. (2011). Natural language processing (almost) from scratch. J Mach Learn Res, 12: 2493–2537

Comunello, F., Parisi, L., Lauciani, V., Magnoni, F., & Casarotti, E. (2016). Tweeting after an earthquake: user localization and communication patterns during the 2012 Emilia seismic sequence Ann. Geophys., 59 (5) (2016), Article 0537

Confuorto, P., Del Soldato, M., Solari, L., Festa, D., Bianchini, S., Raspini, F., & Casagli, N. (2021) Sentinel-1-based monitoring services at Regional scale in Italy: State of the art and main findings. International Journal of Applied Earth Observation and Geoinformation 102(17):102448. DOI:10.1016/j.jag.2021.102448

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver. 7057–7067

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020) Unsupervised Cross-Lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020

Cresci, S., Tesconi, M., Cimino, A., & Dell'Orletta, F. (2015). A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. Proceedings of the 24th international conference on world wide web, WWW'15 (2015), pp. 1195-1200. http://dx.doi.org/10.1145/2740908.2741722.

Crozier, M.J. (2010). Deciphering the effect of climate change on landslide activity: A review. Geomorphology, 124(3-4), 260-267.

Das, S., Dutta, A., Medina, G., Minjares-Kyle, L., & Elgart, Z. (2019). Extracting patterns from Twitter to promote biking. IATSS Research, 43 (1) (2019), pp. 51-59, 10.1016/j.iatssr.2018.09.002

De Andrade, S.C., De Albuquerque, J.P., Restrepo Estrada, C., Westerholt, R., Rodriguez, C.A.M., Mendiondo, E.M., & Botazzo Delbem, A.C. (2021). The effect of intra-urban mobility flows on the spatial heterogeneity of social media activity: investigating the response to rainfall events, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2021.1957898

De Stefano, A. (2002). I lineamenti geologici e strutturali del territorio lucano. Collane Regione Basilicata, Cultura-Il Territorio, "Conoscere la Basilicata", 6 pp.

Del Soldato, M., Rosi, A., Delli Passeri, L., Cacciamani, C., Catani, F., & Casagli, N. (2021). Ten years of pluviometric analyses in Italy for civil protection purposes. Sci Rep 11, 20302 (2021). https://doi.org/10.1038/s41598-021-99874-w

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dharma, L.S.A., & Winarko, E. (2022). "Classifying Natural Disaster Tweet using a Convolutional Neural Network and BERT Embedding," *2022 2nd International Conference on Information Technology and Education (ICIT&E)*, 2022, pp. 23-30, doi: 10.1109/ICITE54466.2022.9759860.

Dong, L., Yang, N., Wang, W., et al. (2019). Unified language model pre-training for natural language understanding and generation. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver. 13042–13054

Dong, Z., Yao, Z., Gholami, A., et al. (2019). Hawq: Hessian aware quantization of neural networks with mixed-precision. In: Proceedings of the International Conference on Computer Vision. Seoul. 293–302

Dragović, N., Vasiljević, Đ., Stankov, U., & Vujičić, M. (2019). Go social for your own safety! Review of social networks use on natural disasters – case studies from worldwide. Open Geosciences. 2019;11(1): 352-366. https://doi.org/10.1515/geo-2019-0028

Du, S., Gu, H., Wen, J., Chen. K., & Van Rompaey, A. (2015). Detecting flood variations in Shanghai over 1949–2009 with Mann-Kendall tests and a newspaper-based database. Water, 7(5), 1808-1824.

Dufty, N. (2012). Using social media to build community disaster resilience. Australian Journal of Emergency Management 27 (1):40.

Earle, P. (2010). Earthquake twitter. Nature Geoscience, 2010

Easterling, D.R., Meehl, G.A., Parmesan, C., Changnon, S.A., Karl, T.R., & Mearns, L.O. (2000). Climate extremes: observations, modelling, and impacts. Science 289:2068–2074. https://doi.org/10.1126/science.289.5487

Ehnis, C., & Bunker, D. (2012). Social media in disaster response: Queensland Police Service-public engagement during the 2011 floods. In: Lamp J. (Ed.), ACIS 2012: Location, location, location. Proceedings of the 23rd Australasian Conference on information systems, Geelong, Australia.

Erhan, D., Bengio, Y., Courville, A.C., et al. (2010). Why does unsupervised pretraining help deep learning? J Mach Learn Res, 11: 625–660

Ester, M., Kriegel, H-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in a large spatial databases with noise. Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, pp. 226-231.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1998). Clustering for mining in large spatial databases. KI-Journal (Artificial Intelligence), Special Issue on Data Mining. 12(1). 18-24.

Faccini, F., Luino, F., Sacchini, A., & Turconi, L. (2015). Flash flood events and urban development in Genoa (Italy): lost in translation. In Engineering Geology for Society and Territory-Volume 5 (pp. 797-801). Springer, Cham.

Fan, C., Mostafavi, A., Gupta, A., & Zhang, C. (2018). A system analytics framework for detecting infrastructure-related topics in disasters using social sensing in Advanced Computing Strategies for Engineering, Cham, Switzerland: Springer, pp. 74-91.

Fan, C., Jiang, Y., & Mostafavi, A. (2020). Social sensing in disaster city digital twin: Integrated textual–visual–geo framework for situational awareness during built environment disruptions. Journal of Management in Engineering, 36(3), 04020002.

Fayjaloun, R., Gehl, P., Auclair, S., Boulahya, F., Guérin-Marthe, S., & Roullé, A. (2020). Integrating strong-motion recordings and Twitter data for a rapid shakemap of macroseismic intensity, Int. J. Disast. Risk Reduct. 52, 101927, 2212–4209, doi: 10.1016/ j.ijdrr.2020.101927.

Fitriany, A.A., Flatau, P.J., Khoirunurrofik, K. & Riama, N.F. (2021). Assessment on the Use of Meteorological and Social Media Information for Forest Fire Detection and Prediction in Riau, Indonesia. Sustainability 2021, 13, 11188. https://doi.org/10.3390/su132011188

Fischer, H.W. (1994). Response to Disaster: Fact versus Fiction and Its Perpetuation. University Press of America (160 pp.)

Forli, A., & Guida, T. (2009). Il rischio idrogeologico in Italia. Adempimenti e tecniche operative d'intervento. Sistemi Editoriali, Roma, 412 p. ISBN 9788851305727

Francalanci, C., Guglielmino, P., Montalcini, M., Scalia, G., & Pernici, B. (2017). IMEXT: a method and system to extract geolocated images from Tweets—analysis of a case study. 2017 11th International Conference on Research Challenges in Information Science (RCIS), IEEE (2017), pp. 382-390

Franceschini, R., Rosi, A., Catani, F., & Casagli, N. (2022a). Exploring a landslide inventory created by automated web data mining: the case of Italy. Landslides, 1-13.

Franceschini, R., Rosi, A., Del Soldato, M., Catani, F., & Casagli, N. (2022b). Integrating multiple information sources for landslide hazard assessment: the case of Italy. Scientific Reports, 12(1), 20724.

Fraustino, J.D., Brooke, L., & Yan, J. (2012). Social Media Use during Disasters: A Review of the Knowledge Base and Gaps. Final Report to Human Factors/Behavioural Sciences Division, Science and Technology Directorate, U.S. Department of Homeland Security, National consortium for the study of terrorism and responses to terrorism (START), College Park, Maryland, 2012

Galli, M., Ardizzone, F., Cardinali, M., Guzzetti, F., & Reichenbach, P. (2008). Comparing landslide inventory maps Geomorphology, 94, pp. 268-289.

Ganesh, P., Chen, Y., Lou, X., Khan, M.A., Yang, Y., Sajjad, H., & Winslett, M. (2021). Compressing large-scale transformer-based models: A case study on bert. Transactions of the Association for Computational Linguistics, 9, 1061-1080.

Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the Crowdsourcing Power of Social Media for Disaster Relief. IEEE Intell. Syst., 2011, 26 (3), 10-14, 10.1109/MIS.2011.52

Gao, L., Song, C., Gao, Z., L. Barab´asi, A., Bagrow, J.P., & Wang, D. (2014). Quantifying information flow during emergencies. Scientific reports, 2014.

Geetha, M.P., & Renuka, D.K. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. International Journal of Intelligent Networks, 2, 64-69.

Gehring, J., Auli, M., Grangier, D., et al. (2017). Convolutional sequence to sequence learning. In: Proceedings of the International Conference on Machine Learning. Sydney, 1243–1252

Giardino, M., Giordan, D., & Ambrogio, S. (2004). G.I.S technologies for data collection, management and visualization of large slope instabilities: two applications in the Western Italian Alps. Nat Hazards Earth SystSci 4:197–199 205

Giordano, L., Giordano, F., Grauso, S., Iannetta, M., Rossi, L., Sciortino, M., & Bonati, G. (2002). Individuazione delle aree sensibili alla desertificazione nella regione siciliana (sensitive areas to desertification in Sicily, Italy). In: Iannetta M, Borrelli G (eds) Valutazione e mitigazione della desertificazione nella regione Sicilia: un caso di studio. Enea, Roma, pp 27–47

Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A., & Chakraborty, B. (2018). A review on application of data mining techniques to combat natural disasters. Ain Shams Eng. J., 2018, 9, 365–378, 10.1016/j.asej.2016.01.012

Grauso, S., Pagano, A., Fattoruso, G., et al. (2008). Relations between climatic–geomorphological parameters and sediment yield in a mediterranean semi-arid area (Sicily, southern Italy). Environ Geol 54, 219–234. https://doi.org/10.1007/s00254-007-0809-4

Gründer-Fahrer, S., Schlaf, A., Wiedemann, G., & Heyer, G. (2018). Topics and topical phases in German social media communication during a disaster. Natural language engineering, 24(2), 221–264. https://doi.org/10.1017/S1351324918000025.

Guan, J., Huang, F., Zhao, Z., Zhu, X., & Huang, M. (2020). A knowledge-enhanced pretraining model for commonsense story generation. Transactions of the Association for Computational Linguistics, 8, 93-108.

Guo, Q., Qiu, X., Liu, P., et al. (2019). Star-transformer. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis. 1315–1325

Guzzetti, F. (2000). Landslide fatalities and evaluation of landslide risk in Italy. Engineering Geology 58:89–107.

Hamon, T., Fraisse, A., Paroubek, P., Zweigenbaum, P., & Grouin, C. (2015). Analyse des ´emotions, sentiments et opinions exprim´es dans les tweets : pr´esentation et r´esultats de l ' ´edition 2015 du d´efi fouille de texte (DEFT). Actes de la 22e conference sur le Traitement Automatique des Langues Naturelle ´ s.

Hinton, G.E. (2006). Reducing the dimensionality of data with neural networks. Science, 313: 504–507

Hinton, G., McClelland, J., & Rumelhart, D. (1990). Distributed representations. The Philosophy of Artificial Intelligence, 248–280

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2(7).

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S. & Kanae, S. (2013). Global flood risk under climate change. Nature climate change, 3(9), 816-821.

Holderness, T., & Turpin, E. (2015). From social media to geosocial intelligence: crowdsourcing civic co-management for flood response in Jakarta, Indonesia. In Social Media for Government Services, 115–133 (Springer, 2015).

Horita, F.E., de Albuquerque, J.P., Marchezini, V., & Mendiondo, E.M. (2017). Bridging the gap between decision-making and emerging big data sources: an application of a model-based framework to disaster management in Brazil Decis. Support Syst., 97, pp. 12-22, 10.1016/j.dss.2017.03.001 http://www.sciencedirect.com/science/article/pii/S0167923617300416

Houston, J.B., Hawthorne, J., Perreault, M.F., Park, E.H., Goldstein-Hode, M., Halliwell, M.R. et al. (2015). Social media and disasters: a functional framework for social media use in disaster planning, response, and research. Disasters, 2015, 39 (1), 1–22, 10.1111/disa.12092

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Hu, Y., Zhang, Q., Zhao, W., & Wang, H. (2021). TransQuake: A transformer-based deep learning approach for seismic P-wave detection. Earthquake Research Advances, 1(2), 100004.

Huang, K., Altosaar J., & Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. ArXiv: 1904.05342

Huang, L., Shi, P., Zhu, H., & Chen, T. (2022). Early detection of emergency events from social media: A new text clustering approach. Natural Hazards, 111(1), 851-875.

Huang, Y., Li, Y., & Shan, J. (2018). Spatial-temporal event detection from geo-tagged tweets. ISPRS International Journal of Geo-Information, 7(4), 150. https://doi.org/10.3390/ijgi7040150.

Huang, Z., Zhou, J., Song, L., Lu, Y., & Zhang, Y. (2010). Flood disaster loss comprehensive evaluation model based on optimization support vector machine. Expert Systems with Applications, 37(5), 3810-3814.

Huffington Post: How Hurricane Sandy impacted internet usage through Netflix, Skype and more, 2012 http://www.huflngtonpost.com/2012/11/02/how-hurricane-sandy-impac_n_2066515.html

Hughes, A.L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management, 6(3), 248–260.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) 47, 4, 67.

Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. Proceedings of the tenth international conference on language resources and evaluation (LREC'2016) European Language Resources Association (ELRA).

IPCC, 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK, and New York, NY, USA. doi:https://doi.org/10.1017/CBO9781139177245.009.

Irons, M., Paton, D., Lester, L., Scott, J., & Martin, A. (2014). Social media, crisis communication and community-led response and recovery: An Australian case study. Proceedings of the Research Forum at the Bushfires and Natural Hazards CRC & AFAC conference, Wellington, New Zealand, Report No. 2015.056

ISPRA (2020a). Land use, spatial dynamics and ecosystem services. Report2020. (In Italian)

ISPRA (2020b). Climatic indicators of Italy in 2018 (in Italian)

Jain, P., Ross, R., & Schoen-Phelan, B. (2019, August). Estimating distributed representation performance in disaster-related social media classification. In 2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM) (pp. 723-727). IEEE.

Joshi, M., Chen, D., Liu, Y., et al. (2019). SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8: 64–77

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Baltimore, 655–665

Keiler, M., Knight, J., & Harrison, S. (2010). Climate change and geomorphological hazards in the eastern European Alps. Philos Trans R Soc 368:2461–2479. https://doi.org/ 10.1098/rsta.2010. 0

Kersten, J., & Klan, F. (2020). What happens where during disasters ? A Workflow for the multifaceted characterization of crisis events based on Twitter data. J Contingencies and Crisis Management 2020 ;28 :262-280. https://doi.org/10.1111/1468-5973.12321

Kim, Y. (2014). Convolutional neural networks for sentence classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, 1746–1751

Kim, J., & Hastak, M. (2018). Social network analysis: Characteristics of online social networks after a disaster. Int. J. Inform. Manage., 2018, 38 (1), 86-9610.1016/j.ijinfomgt.2017.08.003

Kiely, G. (1999). Climate change in Ireland from precipitation and stream flows observations. Adv. Water Resour. 23, 141–151.

Kirschbaum, D., Stanley, T., & Zhou, Y. (2015). Spatial and temporal analysis of a global landslide catalog. Geomorphology 249:4–15

Klose, M., Damm, B., & Highland, L. (eds) (2015). Geohazard databases: concepts, development, Applications locations [Special Issue]. Geomorphology 249:1–136

Knabb, R.D., Jamie, R.R., & Daniel, P.B. (2005). Tropical Cyclone Report: Hurricane Katrina. National Hurricane Center, Miami.

Knight, J., & Harrison, S. (2009). Sediments and future climate. Nat Geosci 2:230. https://doi.org/10.1038/ngeo49

K, K., Wang, Z., Mayhew, S., et al. (2020). Cross-lingual ability of multilingual BERT: An empirical study. In: Proceedings of the International Conference on Learning Representations. Addis Ababa.

Kozlowski, D., Lannelongue, E., Saudemont, F., Benamara, F., Mari, A., Moriceau, V., & Boumadane, A. (2020). A three-level classification of French tweets in ecological crises. Information Processing and Management 57, 102284. https://doi.org/10.1016/j.ipm.2020.102284

Kreuzer, T.M., & Damm, B. (2020). Automated digital data acquisition for landslide inventories. Landslides 17, pages2205–2215.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity Science advances, 2 (3) (2016), Article e1500779

Kryvasheyeu, Y., Chen, H., Moro, E., Van Hentenryck, P., & Cebrian, M. (2015). Performance of social network sensors during Hurricane Sandy. PLoS one, 10(2), e0117288.

Kundzewicz, Z.W., Kanae, S., Seneviratne, S.I., Handmer, J., Nicholls, N., Peduzzi, P. & Sherstyukov, B. (2014). Flood risk and climate change: global and Regional perspectives. Hydrological Sciences Journal, 59(1), 1-28.

Lagomarsino, D., Segoni, S., Fanti, R., & Catani, F. (2013). Updating and tuning a regional-scale landslide early warning system. Landslides, 10(1), 91-97.

Lample, G., & Conneau, A. (2019). Cross-Lingual Language Model Pretraining. arXiv 2019, arXiv:1901.07291.

Legambiente (2021). Osservatorio Città Clima: https://cittaclima.it/2021/12/29/emergenza-clima-il-bilancio-del-2021-dellosservatorio-cittaclima/

Levine, Y., Lenz, B., Dagan, O., Ram, O., Padnos, D., Sharir, O., & Shoham, Y. (2019). Sensebert: Driving some sense into bert. arXiv preprint arXiv:1908.05646.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., & Schwab, D. (2019). Flaubert: Unsupervised language model pre-training for french. arXiv preprint arXiv:1912.05372.

Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36: 1234–1240

Lee, E., Lee, C., & Ahn, S. (2022). Comparative Study of Multiclass Text Classification in Research Proposals Using Pretrained Language Models. Applied Sciences, 12(9), 4522.

Leykin, D., Lahad, M., & Aharonson-Daniel, L. (2018). Gauging urban resilience from social media. International journal of disaster risk reduction, 31, 393-402.

Li, L., Bensi, M., Cui, Q., Baecher, G.B., & Huang, Y. (2021). Social media crowdsourcing for rapid damage assessment following a sudden-onset natural hazard event. Int. J. Inf. Manag., v60 (2021), p. 102378, 10.1016/j.ijinfomgt.2021.102378

Liddy, E.D. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

Liu, J., Singhal, T., Blessing, L.T., Wood, K.L., & Lim, K.H. (2021, August). Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media (pp. 133-141).

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. In: Proceedings of the International Joint Conference on Artificial Intelligence. New York. 2873– 2879

Liu, W., Zhou, P., Zhao, Z., et al. (2020). K-BERT: Enabling language representation with knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. New York. 2901–2908

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D. & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Lobb, A., Mock, N., & Hutchinson, P.L. (2012). Traditional and social media coverage and charitable giving following the 2010 earthquake in Haiti. Prehospital and Disaster Medicine, 2012, 27 (4), 319–324, 10.1017/S1049023X12000908

Loayza, N.V., Olaberria, E., Rigolini, J., & Christiaensen, L. (2012). Natural disasters and growth: Going beyond the averages. World Development, 40(7), 1317–1336.

Lu, P., Casagli, N., Catani, F., & Tofani, V. (2012) Persistent scatterers interferometry hotspot and cluster analysis (PSI-HCA) for detection of extremely slow-moving landslides. Int J Remote Sens 33(2):466e489

Madichetty, S., & Muthukumarasamy, S. (2020). Detection of situational information from Twitter during disaster using deep learning models. Sādhanā, 45(1), 1-13.

Madichetty, S., & Sridevi, M. (2020). Improved classification of crisis-related data on Twitter using contextual representations. Procedia Comput. Sci., vol. 167, pp. 962-968.

Magro, M.J. (2012). A Review of Social Media Use in E-Government. Administrative Sciences, 2012, 2, 148-161, 10.3390/admsci2020148

Majumdar, A., & Bose, I. (2019). Do tweets create value? A multi-period analysis of Twitter use and content of tweets for manufacturing firms. International Journal of Production Economics, 216 (2019), pp. 1-11, 10.1016/j.ijpe.2019.04.008

Mahoney, J., Le Moignan, E., Long, K., Wilson, M., Barnett, J., Vines, J., & Lawson, S. (2019). Feeling alone among 317 million others: Disclosures of loneliness on Twitter. Computers in Human Behavior, 98 (2019), pp. 20-30, 10.1016/j.chb.2019.03.024

Marcheggiani, D., Bastings, J., & Titov, I. (2018). Exploiting semantics in neural machine translation with graph convolutional networks. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans. 486–492

Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de La Clergerie, É.V., & Sagot, B. (2019). CamemBERT: a tasty French language model. arXiv preprint arXiv:1911.03894.

Mashable: Hurricane Sandy is 2012's No. 2 Topic on Facebook, 2012 http://mashable.com/2012/10/31/hurricane-sandy-facebook/

Mav Social: How Social Media Helped during Typhoon Haiyan, 2013 https://mavsocial.com/how-social-media-helped-during-typhoon-haiyan/

McCreadie, R., & Soboroff, I. (2018). TREC overview paper – incidents stream track. In 26th Text REtrieval Conference (TREC), 14th November (p.9).

McKean, J., & Roering, J. (2003) Objective landslide detection and surface morphology mapping using high-resolution airborne laser altimetry.Geomorphology 57(3e4):331e351

Mikolov, T., Sutskever, I., Chen, K., et al. (2013). Distributed representations of words and phrases and their compositionality. In: Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe. 3111–3119

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In: Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. Atlanta. 746–751

Miles, B., & Morse, S. (2007). The role of news media in natural disaster risk and recovery. Ecological economics, 63(2-3), 365-373.

Mills, A., Chen, R., Lee, J., & Rao, H.R. (2009) Web 2.0 emergency applications: How useful can Twitter be for an emergency response? Journal of Information Privacy and Security, 5 (3), 3-26

Miner, G., Elder, I.V.J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Montalti, R., Solari, L., Bianchini, S., Del Soldato, M., Raspini, F., & Casagli, N. (2019). A Sentinel-1-based clustering analysis for geo-hazards mitigation at Regional scale: a case study in Central Italy. Geomatics, Natural Hazards and Risk, 10(1), 2257-2275.

Murias, P., Novello, S., & Martinez, F. (2012). The regions of economic well-being in Italy and Spain. Regional Studies, 46(6), 793-816. DOI: 10.1080/00343404.2010.504702

Musaev, A., Wang, D., & Pu, C. (2014). LITMUS: a multi-service composition system for landslide detection. IEEE Transactions on Services Computing, 8(5), 715-726.

National Academies of Sciences & Medicine. Framing the Challenge of Urban Flooding in the United States. https://www.nap.edu/catalog/25381/framingthe-challenge-of-urban-flooding-in-the-united-states. (The National Academies Press, Washington, DC, 2019)

Nava, L., Bhuyan, K., Meena, S.R., Monserrat, O., & Catani, F. (2022). Rapid Mapping of Landslides on SAR Data by Attention U-Net. Remote Sensing, 14(6), 1449.

Neumayer, E., Plümper, T. & Barthel, F. (2014) The political economy of natural disaster damage. Global Environmental Change Volume 24, January 2014, Pages 8-19. https://doi.org/10.1016/j.gloenvcha.2013.03.011

Nguyen, D.T., Al-Mannai, K.A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. In Proceedings of the 11th international conference on web and social media, ICWSM 2017 (pp. 632-635). AAAI press.

Nugroho, Y. (2011). Citizens in @action: Collaboration, participatory democracy and freedom of information – Mapping contemporary civic activism and the use of new social media in Indonesia. Research collaboration of Manchester Institute of Innovation Research, University of Manchester and HIVOS Regional Office Southeast Asia, Manchester, United Kingdom, Jakarta, Indonesia.

Olteanu, A., Vieweg, S., & Castillo, C. (2015). 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, New York, NY, 994–1009. DOI:http://dx.doi.org/10.1145/2675133.2675242

Osorio-Arjona, J., & García-Palomares, J.C. (2019). Social media and urban mobility: Using twitter to calculate home-work travel matrices. Cities, 89 (2019), pp. 268-280, 10.1016/j.cities.2019.03.006

Padli, J., Habibullah, M.S., & Baharom, A.H. (2018). The impact of human development on natural disaster fatalities and damage: panel data evidence. Pages 1557-1573. DOI: https://doi.org/10.1080/1331677X.2018.1504689

Paliaga, G., Luino, F., Turconi, L., Marincioni, F., & Faccini, F. (2020) Exposure to Geo-Hydrological Hazards of the Metropolitan Area of Gennoa, Italy: A Multi-Temporal Analysis of the Bisagno Stream. Sustainability 2020, 12(3), 1114. https://doi.org/10.3390/su12031114

Panizza, M., Corsini, A., Ghinoi, A., Marchetti, M., Pasuto, Al., & Soldati, M. (2011). Explanatory notes of the geomorphological map of the alta Badia Valley (Dolomites, Italy). Geogr. Fis. Dinam. Quat. 34 (2011), 105-126.

Park, J., Seager, T. P., Rao, P. S. C., Convertino, M., & Linkov, I. (2013). Integrating risk and resilience approaches to catastrophe management in engineering systems. Risk analysis, 33(3), 356-367.

Peters, M.E., Neumann, M., IV, R. L. L., et al. (2019). Knowledge enhanced contextual word representations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Hong Kong. 43–54

Pirnau, M. (2017). Analysis of data volumes circulating in SNs after the occurrence of an earthquake. Rom. J. Inf. Sci. Tech., 20 (3), 286-298

Plunz, R.A., Zhou, Y., Carrasco Vintimilla, M.I., Mckeown, K., Yu, T., Uguccioni, L., & Sutto, M.P. (2019). Twitter sentiment in New York City parks as measure of well-being. Landscape and Urban Planning, 189 (2019), pp. 235-246, 10.1016/j.landurbplan.2019.04.024

Polaris (2014). Rapporto periodico sul rischio posto alla popolazione da frane e inondazioni. Consiglio Nazionale delle Ricerche. Istituto di ricerca per la protezione idrogeologica (IRPI) Published: 2015

Polaris (2019). Rapporto periodico sul rischio posto alla popolazione da frane e inondazioni. Consiglio Nazionale delle Ricerche. Istituto di ricerca per la protezione idrogeologica (IRPI) Published: 2020

Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., & Basile, V. (2019). Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In 6th Italian Conference on Computational Linguistics, CLiC-it 2019 (Vol. 2481, pp. 1-6). CEUR.

Porfiriev, B.N. (2016). The economics of natural disasters. ISSN 10193316, Herald of the Russian Academy of Sciences, 2016, Vol. 86, No. 1, pp. 1–11. DOI: 10.1134/S1019331616010020

Qadir, J., Ali, A., Rasool, R., Zwitter, A., Sathiaseelan, A., & Crowcroft, J. (2016). Crisis analytics: Big data-driven crisis response. Journal of International Humanitarian Action, 1, 1–21.

Qiu, X.P., Sun, T.X., Xu, Y.G., et al. (2020). Pre-trained models for natural language processing: A survey. Sci China Tech Sci, 63: 1872–1897, https://doi.org/10.1007/s11431-020-1647-3

Rachunok, B.A., Bennett, J.B., & Nateghi, R. (2019). Twitter and disasters: A social resilience fingerprint, *IEEE Access*, vol. 7, pp. 58495-58506.

Rachunok, B., Bennett, J., Flage, R., & Nateghi, R. (2021). A path forward for leveraging social media to improve the study of community resilience. International Journal of Disaster Risk Reduction, 59, 102236.

Radford, A., Narasimhan, K., Salimans, T., et al. (2018). Improving language understanding by generative pre-training. OpenAI Blog.

Ramachandran, P., Liu, P.J., & Le, Q. (2017). Unsupervised pretraining for sequence to sequence learning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen. 383–391

Ratto, S., Giardino, M., Giordan, D., Alberto, W., & Armand, M. (2007). Carta dei fenomeni franosi della Valle d'Aosta, scala 1: 100.000-Regione Autonoma Valle d'Aosta, Assessorato Territorio, Ambiente e Opere Pubbliche.

Rebetez, M., Lugon, R., & Baeriswyl, P.A. (1997). Climatic change and debris flows in high mountain regions: the case study of the Ritigraben torrent (Swiss Alps). Clim Chang 36:371–389. https://doi.org/10.1023/A:1005356130392

Reboredo, J.C., & Ugolini, A. (2018). The impact of Twitter sentiment on renewable energy stocks. Energy Economics, 76 (2018), pp. 153-169, 10.1016/j.eneco.2018.10.014

Reich, J. W., Zautra, A. J., & Hall, J. S. (Eds.). (2010). Handbook of adult resilience. Guilford Press.

Reuter, C., Lee-Hughes, A., & Kaufhold, M.A. (2018). Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. Int. J. Hum–Comput. Int., 2018, 34 (4), 280-294, 10.1080/10447318.2018.1427832

Romascanu, A., Ker, H., Sieber, R., Greenidge, S., Lumley, S., Bush, D., Morgan, S., Zhao, S., & Brunila, M. (2020). Using deep learning and social network analysis to understand and manage extreme flooding. J Contingencies and Crisis Management 28-251-261.

Rosenthal, S., Farra, N., & Nakov, P. (2018). SemEval-2017 Task 4: Sentiment Analysis in Twitter. pages 502–518

Rosenzweig, C., Karoly, D., Vicarelli, M., Neofotis, P., Wu, Q., Casassa, G., Menzel, A., Root, T.L., Estrella, N., Seguin, B., Tyrjanowski, P., Liu, C., Rawlins, S., & Imenson, A. (2008). Attributing physical and biological impacts to anthropogenic climate change. Nature 453:353–357. https://doi.org/10.1038/nature069

Rosi, A., Tofani, V., Tanteri, L., Tacconi Stefanelli, C., Agostini, A., Catani, F., & Casagli, N. (2018). The new landslide inventory of Tuscany (Italy) updated with PS-InSAR: geomorphological features and landslide distribution. Landslides 15:5–19. https://doi.org/10.1007/s10346-017-086

Rosi, A., Segoni, S., Canavesi, V., Monni, A., Gallucci, A., & Casagli, N. (2021). Definition of 3D rainfall thresholds to increase operative landslide early warning system performances. Landslides 18(3):1045–1057

Saltelli, A., Bammer, G., Bruno, I., Charters, E., Di Fiore, M., Didier, E., Espeland, W.N., Kay, J., Lo Piano, S., Mayo, D., Pielke, R. Jr., Portaluri, T., Porter, T.M., Puy, A., Rafols, I., Ravets, J.R., Reinert, E., Sarewitz, D., Stark, P.B., et al. (2020). Five ways to ensure that models serve society: a manifesto. Nature 582:482–483. https://doi.org/10.1038/d41586-020-01812-9

Salvati, P., Bianchi, C., Rossi, M., & Guzzetti, F. (2010). Societal landslide and flood risk in Italy. Nat. Hazards Earth Syst. Sci. 2010, 10, 465–483.

Salvatici, T., Tofani, V., Rossi, G., D'Ambrosio, M., Tacconi Stefanelli, C., Masi, E.B., Rosi, A., Pazzi, V., Vannocci, P., Petrolo, M., Catani, F., Ratto, S., Stevenin, H., & Casagli, N. (2018). Application of a physically based model to forecast shallow landslides at a Regional scale. Nat Hazards Earth Syst Sci 18:1919–1935. https://doi.org/10.5194/nhess- 18- 1919- 20

Sánchez, C., Sarmiento, H., Pérez, J., Abeliuk, A., & Poblete, B. (2022). Cross-Lingual and Cross-Domain Crisis Classification for Low-Resource Scenarios. arXiv preprint arXiv:2209.02139.

Santangelo, M., Cardinali, M., Rossi, M., Mondini. A.C., & Guzzetti, F. (2010). Remote landslide mapping using a laser rangefinder binocular and GPS. Natural Hazards and Earth System Science, 10, pp. 2539-2546.

Saunshi, N., Plevrakis, O., Arora, S., et al. (2019). A theoretical analysis of contrastive unsupervised representation learning. In: Proceedings of the International Conference on Machine Learning. Long Beach, 2019. 5628–5637

Segoni, S., Tofani, V., Rosi, A., Catani, F., & Casagli, N. (2018) Combination of Rainfall Thresholds and Susceptibility Maps for Dynamic Landslide Hazard Assessment at Regional Scale. Front. Earth Sci. 6:85. Doi:10.3389/feart.2018.00085

Shoyama, K., Cui, Q., Hanashima, M., Sano, & Usada, Y. (2021). Emergency flood detection using multiple information sources: Integrated analysis of natural hazard monitoring and social media data. Science of The Total Environment, Volume 767, 1 May 2021, 144371. https://doi.org/10.1016/j.scitotenv.2020.144371

Singh, J.P., Dwivedi, Y.K., Rana, N.P., Kumar, A., & Kapoor, K.K. (2019). Event classification and location prediction from tweets during disasters. Annals of Operations Research, 283(1), 737-757.

Slide Share: Metrics report: Special edition: Hurricane Sandy, 2012 https://www.slideshare.net/bonagreg/hurricane-sandy-websocial-metrics-report

Socher, R., Perelygin, A., Wu, J.Y., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Seattle. 1631–1642

Soeters, R., & Van Westen, C. (1996). Slope instability recognition, analysis and zonation, in Landslides, investigation and mitigation. TransportationResearch Board, National Research Council, National Academy Press,Washington, p 129e177

Solari, L., Bianchini, S., Franceschini, R., Barra, A., Monserrat, O., Thuegaz, P., Bertolo, D., Crosetto, M., & Catani, F. (2020) Satellite interferometric data for landslide intensity evaluation in mountainous Regions. Int J Appl Earth Obs Geoinf 87:102028. https://doi.org/10.1016/j.jag.2019.1020

Spruce, M., Arthur, R., & Williams, H.T.P. (2020). Using social media to measure impacts of named storm events in the United Kingdom and Ireland. Meteorological Applications, 27 (1) (2020), 10.1002/met.1887

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, October). How to fine-tune bert for text classification?. In China national conference on Chinese computational linguistics (pp. 194-206). Springer, Cham.

Sutskever, I., Vinyals, O., & Le, Q.V. (2014). Sequence to sequence learning with neural networks. In: Proceedings of the Advances in Neural Information Processing Systems. Montreal. 3104–3112

Tai, K.S., Socher, R., & Manning, C.D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Beijing. 1556–1566

Taylor, W.L. (1953). “Cloze Procedure”: A new tool for measuring readability. Jism Q, 30: 415–433

Teodorescu, H.N. (2015). Using analytics and social media for monitoring and mitigation of social disasters. Procedia Engineering, 107, 325 – 334,

The World Bank and The United Nations. (2010). Natural hazards, unnatural disasters: The economics of effective prevention. Washington DC: The World Bank.

TIZ - Active Twitter users: https://www.tiz.fr/utilisateurs-reseaux-sociaux-france-monde/, last access in July 2020.

Trigila, A., Iadanza, C., & Guerrieri, L. (2007). The IFFI project (Italian landslide inventory): Methodology and results. Guidelines for Mapping Areas at Risk of Landslides in Europe, edited by: Hervás, J., ISPRA, Rome, Italy, 15-18.

Trigila, A., & Iadanza, C. (2018). Landslides and floods in Italy: hazard and risk indicators-Summary Report 2018. Report number: 267bis/2018. Affiliation: Institute for Environmental Protection and Research (ISPRA). Project: National risk indicators. DOI:10.13140/RG.2.2.14114.48328

UN World Conference on Disaster Risk Reduction. (March 14–18, 2015). Sendai, Japan. http:// www.preventionweb.net/files/ 45069_ proceedingsthir dunitednationsworldc.pdf. Cited January 2, 2016.

Vai, F., & Martini, I.P. (2001). Anatomy of an Orogen: the Apennines and adjacent Mediterranean basins, XVIII edn. Springer Netherlands, Netherlands, p 633. https://doi.org/10.1007/978-94-015-982

Van Den Eeckhaut, M., & Hervás, J. (2011). State of the art of national landslide databases in Europe and their potential for assessing landslide susceptibility, hazard and risk. Geomorphology (2011), 10.1016/j.geomorph.2011.12.006

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems. Long Beach. 5998–6008

Verma, S., Vieweg, S., Corvey, W.J., Palen, L., Martin, J.H., Palmer, M., Schram, A., & M Anderson, K. (2011). Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, No. 1, pp. 385-392).

Vieweg, S. (2012). Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications. Ph.D. Dissertation. University of Colorado at Boulder.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076.

Wang, R.Q., Mao, H., Wang, Y., Rae, C., & Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. Computers Geosci. 111, 139–147.

Wang, K., Lam, N.S., Zou, L., & Mihunov, V. (2021). Twitter use in hurricane isaac and its implications for disaster resilience. ISPRS International Journal of Geo-Information, 10(3), 116.

Wang, K., Lam, N.S., & Mihunov, V. (2023). Correlating Twitter Use with Disaster Resilience at Two Spatial Scales: A Case Study of Hurricane Sandy. Annals of GIS, 1-20.

Xiong, W., Du, J., Wang, W.Y., et al. (2020). Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In: Proceedings of the International Conference on Learning Representations. Addis Ababa.

Yang, Z., Dai, Z., Yang, Y., et al. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver. 5754–5764

Zhang, X. A., & Shay, R. (2019). An examination of antecedents to perceived community resilience in disaster postcrisis communication. Journalism & Mass Communication Quarterly, 96(1), 264–287.

Zhang, Z., Zou, Y., & Gan, C. (2018). Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression. Neurocomputing 275:1407–1415. https://doi.org/10.1016/j.neucom.2017.09.080

Zhong, X., Duckham, M., Chong, D., & Tolhurst, K. (2016). Real-time estimation of wildfire perimeters from curated crowdsourcing. Sci. Rep. 6:24206. doi: 10.1038/srep24206

Zhou, C., Lee, C.F., Li, J., & Xu, Z.W. (2002). On the spatial relationship between landslides and causative factors on Lantau Island, Hong Kong. Geomorphology, 43(3-4), 197-207

Zhou, X., & Xu, C. (2017). Tracing the spatial-temporal evolution of events based on social media data. ISPRS International Journal of Geo-Information,6(3),88. https://doi.org/10.3390/ijgi6030088

Zhou, A., Zhou, S., Cao, J., Fan, Y., & Hu, Y. (2000). Approaches for scaling DBSCAN algorithm to large spatial databases. Journal of computer science and technology, vol. 15, no. 6, pp. 509–526.

Zhou, B., Zou, L., Mostafavi, A., Lin, B., Yang, M., Gharaibeh, N., & Mandal, D. (2022). VictimFinder: Harvesting rescue requests in disaster response from social media with BERT. Computers, Environment and Urban Systems, 95, 101824.

Web

https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

https://huggingface.co/course/chapter0?fw=pt

https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

http://www.rendis.isprambiente.it/rendisweb/vistepub.jsp

https://developer.twitter.com/en/portal/projects/1464429008729247748/apps

https://github.com/jdfoote/Intro-to-Programming-and-Data-Science/blob/fall2021/extra_topics/twitter_v2_example.ipynb

https://www.youtube.com/watch?v=rQEsIs9LERM&list=LL&index=3&t=9s

https://www.mathworks.com/content/dam/mathworks/mathworks-dot-com/company/events/conferences/matlab-computational-finance-conference-nyc/2019/text-analytics-mathworks-a-link.pdf

https://it.mathworks.com/discovery/natural-language-processing.html

https://it.mathworks.com/help/pdf_doc/textanalytics/index.html

Appendix

Translate of some word

anni: years

borsa (economica): financial

causa: cause

causa frana: cause landslide

chiusa: locked

chiusa traffico: traffic locked

crollo: collapse

crollo emotivo: emotional breakdown

crollo nervosa: mental breakdown

crollo ponte: bridge collapse

crollo sonno: sleep breakdown

crollo viadotto: viaduct collapse

dissesto: instabily

dissesto finanziario: financial disaster

dopo: after

dopo crollo: after collapse

dissesto idrogeologico: hydrogeological instability

domani: tomorrow

dissesto manto: instability road surface

frana: landslide

frana addosso: close landslide

frana crotonese: landslide on crotonian

frana minaccia: landslide threat

Genova frana: Genova landslide

governo: government

idrogeologico: Hydrogeological

Italia: Italy

maltempo: bad weather

maltempo frana: bad weather landslide

maltempo genova: bad weather Genova

manto stradale: road surface

morti: victims

minaccia: threat

minaccia palazzi: builing threat

notte: night

nuovo crollo: new collapse

oggi: Today

palazzi: building

persone: people

ponte: bridge

prima: before

prima crollo: before collapse

provincial: provincial

quattro morti: four victims

rischio crollo: risk collapse

senzo unico: one way

sfollati: displaced people

sicurezza: security

situazione: situation

solo: only

strada: road

strada chiusa: road locked

stradale: road

totale: total

traffico: traffic

traffico frana: traffic landslide

viadotto: viaduct

unico alternato: one way alternate

Acronyms

AI	Artificial intelligence
ANN	Artificial neural network
APIs	Application Programming Interfaces
AUC	Area under the curve
BERT	Bidirectional Encoder Representations from Transformers
BEFILE	Bert for information on landslide events
BRGM	Bureau de Recherches Géologiques et Minières
CLS	Classification
CNN	Convolutional neural networsk
CNR	National Research Council
CV	Computer vision
DL	Deep learning
GPS	Global positioning system
GPT	Generative pretrained transformer
K	Kendall's
IBM	International Business Machines Corporation
IDMs	Injured, Deaths and Missing Injured, Deaths and Missing
IDEMs	Injured, Deaths, Evacuated and Missing
IDW	Inverse distance weight
IFFI	Italian Inventory of Landslides
IRPI	Research Institute for Hydrogeological Protection
ISPRA	Istituto Superiore per la Protezione e la Ricerca Ambientale - Italian Institute for Environmental Protection and Research
LDA	Latent dirichlet allocation
LSTM	Long short-term memory
MIG	Multi-risk Information Gateway
ML	Machine learning
MLM	Masked language model
MM	Million
MT	Machine Translation

NER	Name Entities Recognition
NLP	Natural language processing
NSP	Next sentence prediction
OVVs	Out-of-vocabulary words
PAI	Piano Assetto Idrogeologico
POLARIS	Popolazione a Rischio da Frana e da Inondazione in Italia - Populations at risk from landslides and floods in Italy
R	Pearson coefficent
ReNDIS	National Repository of Soil Defence interventions
RNNs	Recurrent neural networks
ROC	Receiver operating characteristic
S	Spearman
SA	Sentiment analysis
SECAGN	Semantic Engine to Classify and Geotagging News
SEP	Separating segments
SL	Supervised learning
SVM	Support vector machine
TF-IDF	Term frequency–inverse document frequency
TPUs	Tensor processing units
UL	Unsupervised learning
WHZs	Warning hydrological zones
XLM	Cross-lingual language model