SVEUČILIŠTE U SPLITU
FAKULTET ELEKTROTEHNIKE, STROJARSTVA I BRODOGRADNJE

POSLIJEDIPLOMSKI STUDIJ
ELEKTROTEHNIKA I INFORMACIJSKA TEHNOLOGIJA

KVALIFIKACIJSKI ISPIT

Techniques for Leveraging User-Generated Social Media Text in Disaster Management

Selena Knežić Buhovac

Split, studeni 2024.

Contents

1. INTRODUCTION **1**

2. MATERIALS AND METHODS **4**

2.1. Methodology . . . 4

2.2. Search strategy and database sources . . . 4

2.2.1. Inclusion criteria . . . 7

2.2.2. Exclusion criteria . . . 7

3. RELATED REVIEWS **8**

4. NATURAL DISASTERS **17**

5. DATA SOURCES **19**

5.1. Volunteered Geographic Information . . . 19

5.1.1. Social Media . . . 20

5.2. Official reports . . . 22

5.3. Newswire . . . 23

5.4. Social media platform selection for crisis management . . . 23

6. MAIN ANALYTIC TECHNIQUES **25**

6.1. Data extraction . . . 26

6.2. Text processing . . . 27

6.2.1. Text preparation-preprocessing . . . 28

6.2.2. Vectorization . . . 30

6.2.3. Social media posts classification . . . 32

6.3. Social Media Data Integration for Enhanced Data Analysis . . . 35

6.4. Time series . . . 39

6.4.1. Time series prediction models . . . 39

6.4.2. Time series classification . . . 40

7. CONCLUSION **42**

LITERATURE **44**

LIST OF LABELS AND ACRONYMS **52**

ABSTRACT **54**

SAŽETAK **55**

1. INTRODUCTION

Every year, natural disasters have a deep impact on the lives of hundreds of millions of individuals worldwide. These catastrophic events affects human life's, economies, health services, infrastructure, but also pose significant challenges to recovery efforts and necessitate comprehensive strategies for resilience and preparedness.

Social media (SM) has become a crucial channel for rapid information exchange during natural disasters. Users often share images, videos, and information about disasters such as hurricanes, floods, earthquakes, or fires. These data can be useful for managing crisis situations, providing assistance, and rescuing people [1]. It is important to note that information shared on SM must be verified, as these platforms can be susceptible to the spread of fake news or unverified information during crises [2], [3]. During natural disasters, different SM platforms may have varying levels of activity and usage. It can be said that the greatest value of SM data lies in real-time publishing during natural disasters. From such unfiltered data, one can sense the severity, challenges, and needs during a natural disaster.

Various social media platforms are considered valuable sources of information in disaster management. These platforms differ by user profiles and primary focus. Twitter [4], with its focus on short text and offering an API for data retrieval, is widely accepted by researchers. Facebook [5], as the social media platform with the longest tradition, is another commonly used platform. However, the utilization of these platforms varies based on geography, demographics, and the type of disaster.

During natural disasters, these platforms serve as vital tools for information dissemination, aiding in situational awareness and enabling individuals to request assistance, thereby improving disaster response efforts. They also provide a means of communication to reach marginalized communities through local residents' or volunteers' accounts. However, the reliability of such data varies depending on demographic factors like age, gender, race, and education, potentially introducing biases in data analysis. In the paper [6], the authors noted that younger users in urban areas more frequently geotag their posts. Facebook limits accessibility due to privacy issues, and the authors selected Twitter because of its available API.

Text from SM gives contextually more extensive data.People who share posts on social media tend to provide more detailed descriptions of the location and conditions during natural hazards compared to other types of information, such as remote sensing or video surveillance. SM enables users to generate huge amounts of data from ground, making it valuable for academia but also for business intelligence [7]. However, managing this volume poses challenges in ensuring accuracy, truthfulness, and security. Social networking platforms, fueled by smartphone usage, receive and transmit significant data, promptly stored but often outpacing processing capabilities, especially during emergencies. This real-time analysis presents a major challenge for researchers and engineers. Data about natural disaster can be extracted and processed from multiple data sources, but in this paper focus will be on SM textual post processing, as can be seen in Figure 1.1.

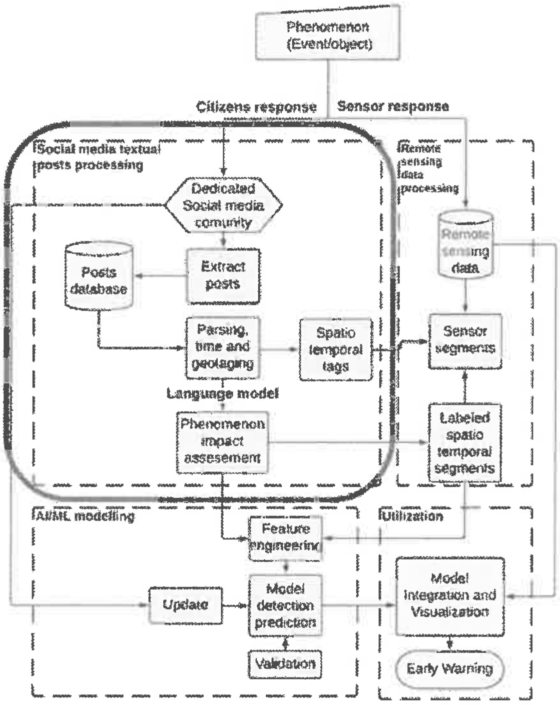
Phenomenon (Event/object)
Citizens response
Sensor response
Social media textual posts processing
Dedicated Social media comunity
Posts database
Extract posts
Parsing, time and geolaging
Spatio temporal tags
Language model
Phenomenon impact assesement
Remote sensing data processing
Remote sensing data
Sensor segments
Labeled spatio temporal segments
AI/ML modelling
Feature engineering
Update
Model detection prediction
Validation
Utilization
Model integration and Visualization
Early Warning

Figure 1.1: Framework for collaborative modelling-main focus [8]

To derive insights from SM posts, various In this paper we distinguish between: data retreaval, preprocessing and filtering, data vectorization, high level inference and time series analysis.

The main aim of this paper is to systematically review the available scientific literature that addresses techniques and algorithms for leveraging user-generated textual data published on social media during natural disasters to improve disaster management. More specifically, the following research questions are posed:

RQ1: What crisis situation-natural and other disasters are considered by SM? What types of disasters are mostly covered in literature and in what extent?

RQ2: What approaches are taken to code user generated data into quantities representation (impact assessment)

RQ3: What techiques are used to transform the user generated data into predictions?

This paper is structured as follows: Section 2 describes the databases used in the study, the search strategies applied, and the results obtained. In Section 3, related reviews and surveys relevant to this research are discussed.

In Sections 4 the most common natural hazards found in the literature are reviewed. Section 5 identifies sources of valuable information about natural hazards. Section 6 discusess difirent types of data and information extraction, preprocessing, vectorization techniques, final processing techniques such as classification and timeseries analysis.

2. MATERIALS AND METHODS

In order to address the posed research questions, a systematic review of the scientific literature was conducted. The methodological approach used in this paper is described in this section.

2.1. Methodology

PRISMA ("Preferred Reporting Items for Systematic Reviews and Meta-Analyses") methodology [9] was applied in this research. When PRISMA methodology is used, four main steps need to be carried out. The first step is 'Planning'. It's focused on research questions and search strategy. Second step is 'selection' which is focused on sorting and extrapolating retrieved data. The third step is 'Extraction'. It's a phase of research intended for evaluating content after applying rigorous criteria for the evaluation. The last step is 'data synthesis'. In this step, the data is analyzed through step-by-step approaches to produce a conclusion [10]. Figure 2.1 illustrates the PRISMA flow diagram used in this research.

2.2. Search strategy and database sources

The Web of Science and Scopus database were used in this research procedure to gain insight into relevant and high-quality information. It is a comprehensive resource for academic and scientific research. These databases are a valuable resource for researchers, academics, and institutions seeking to discover, access, and evaluate scholarly literature and track research trends over time.The search was conducted using the Topic (title, abstract, and keywords) feature. Other databases were excluded to focus solely on the most relevant, high-quality scientific papers.

The query for literature search in WoS: (TS=(social media) OR TS=(social network) OR TS=(VGI)) AND (TS=(disaster) AND TS=(impact) NOT TS=(risk)). WoS categories included: Environmental Sciences, Computer science information systems, Computer science interdisciplinary applications, Computer Science Artificial Intelligence, and Computer Science Theory Methods. For these research query, 131 results were obtained. Figure 2.2 shows the distribution of scientific publications across the mentioned categories.

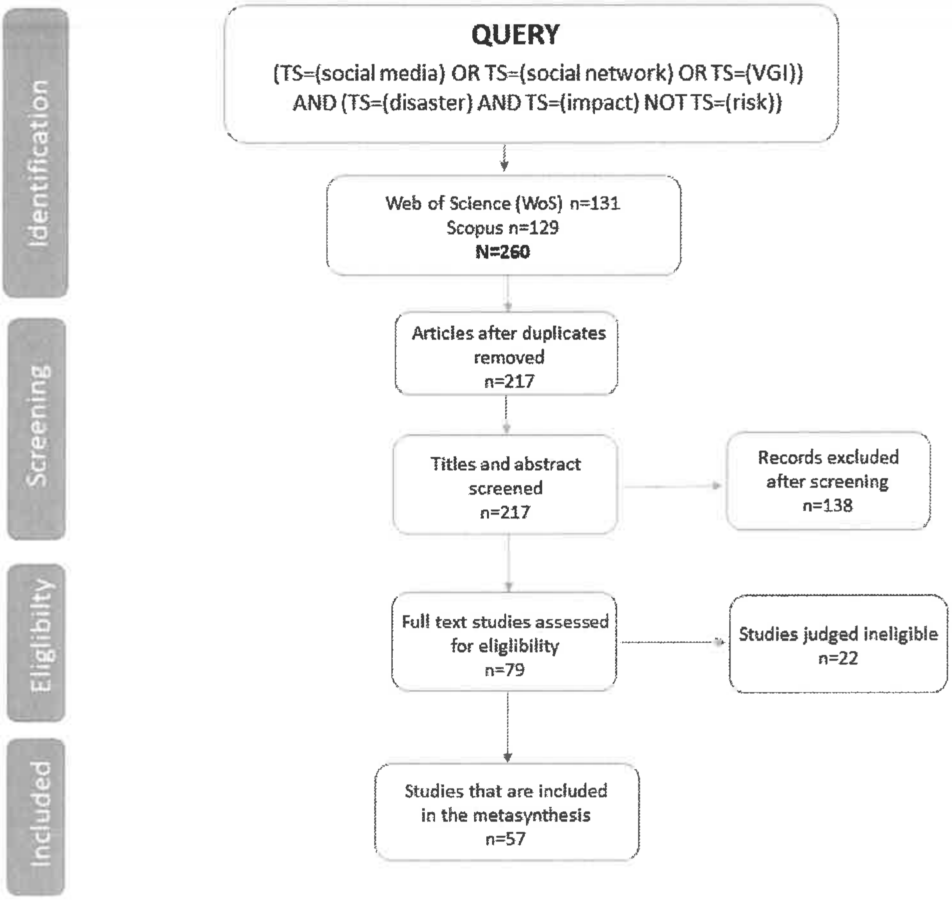
Identification
Screening
Eligibility
Included
QUERY
(TS=(social media) OR TS=(social network) OR TS=(VGI))
AND (TS=(disaster) AND TS=(impact) NOT TS=(risk))
Web of Science (WoS) n=131
Scopus n=129
N=260
Articles after duplicates removed
n=217
Titles and abstract screened
n=217
Records excluded after screening
n=138
Full text studies assessed for eligibility
n=79
Studies judged ineligible
n=22
Studies that are included in the metasynthesis
n=57

Figure 2.1: PRISMA flow diagram (template is re-used and modified from Page et al. [11] with CC BY 4.0)

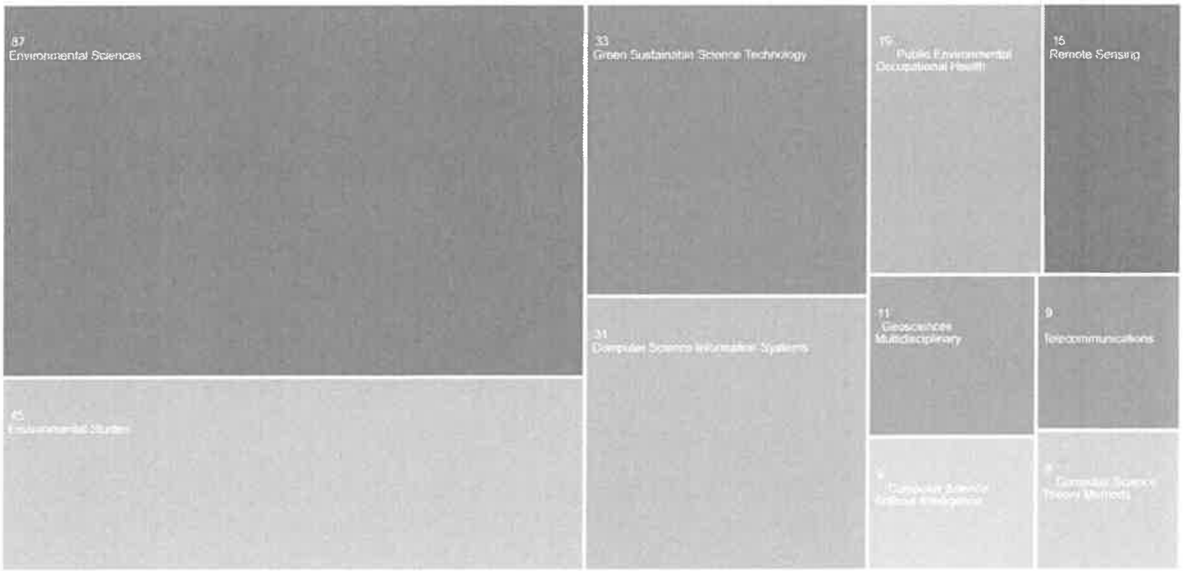

Figure 2.2: Visualization of results by category - WoS

A literature search query in Scopus is similar to a query for the WoS database, but follows the syntax characteristic of the Scopus database. The used query is: (TITLE-ABS-KEY (social AND media) OR TITLE-ABS-KEY (social AND network) OR TITLE-ABS-KEY (vgi) AND TITLE-ABS-KEY (disaster) AND TITLE-ABS-KEY (impact) AND NOT TITLE-ABS-KEY (risk)). The Scopus database has a larger volume of indexed publications compared to WoS, therefore the publications that were obtained using the mentioned query belonged to different categories. The total number of publications was 129, as in the WoS database. The Scopus categories that were the focus of this research are Environmental science and Computer science, which also contained the largest number of publications related to our query. Visualization of all categories and the percentage of publications is illustrated in Figure 2.3.

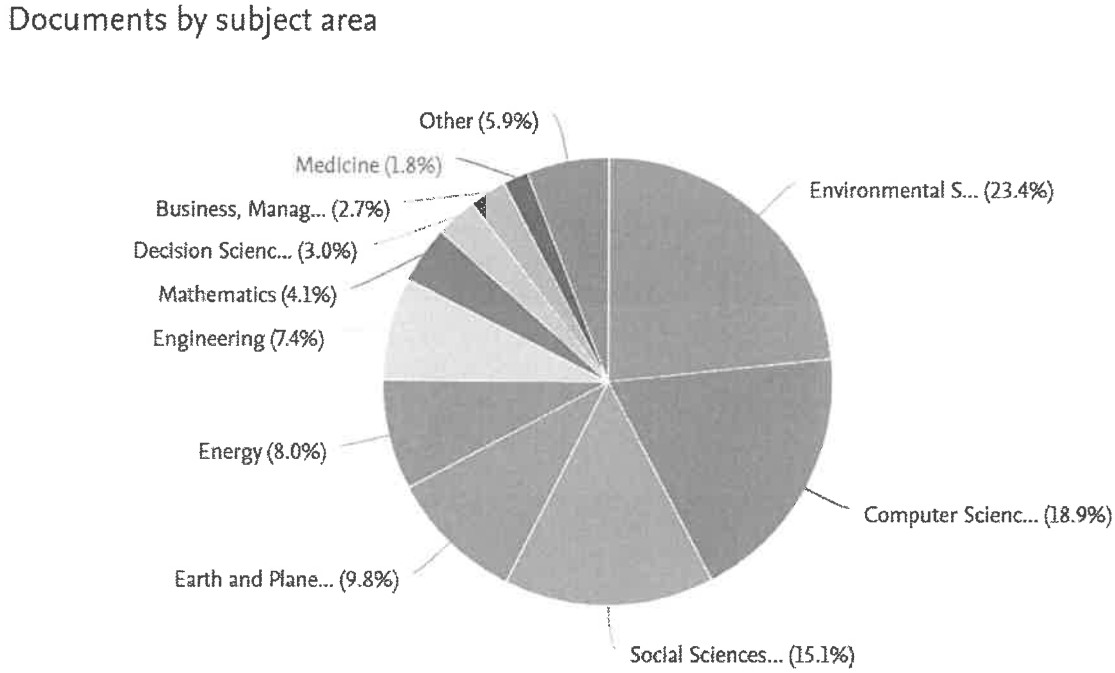

Figure 2.3: Visualization of results by subject area- Scopus

The research used publications published in English in the last five years. Also, the publications are open access for the purpose of a more detailed analysis of them. This is applicable for both databases used in this research.

In order to gain a better understanding of the literature, we had to establish criteria for determining which articles to include in our research. Inclusion and exclusion criteria for articles were set and are described in details in the following sections.

2.2.1. Inclusion criteria

The inclusion criteria defined for this research are as follows:

1. Articles focusing on data and information extraction methodologies, such as machine learning algorithms, natural language processing techniques, or manual coding methods.
2. Articles that include exclusively natural disasters, such as wildfires, floods, hurricanes and earthquakes.
3. Studies on social media data analysis such as time series analysis: Forecasting and prediction, Event Detection and Impact analysis.
4. Evaluation of data extraction techniques.

2.2.2. Exclusion criteria

The exclusion criteria defined for this research are as follows:

1. Articles connected to health services, COVID-19.
2. Articles that analyze the political impact of natural disaster.
3. Articles that analyze social components and emotional consequences during or after natural disaster.
4. Articles about fake news, rumors on social media and their impact on society.
5. Public opinion surveys, survey-based studies.
6. Articles that do not incorporate social media data in their research.
7. Papers related to disaster educations.
8. Image analysis from social media.

From the selected papers included in the analysis, we extracted information on the type of natural disaster, preprocessing techniques, vectorization, final processing such as classification, and the results are presented in the following chapters.

3. RELATED REVIEWS

The literature, consisting of all articles collected from relevant databases, was systematically examined. From the entirety of the literature referenced in the previous section, reviews and surveys were singled out and subjected to a more detailed analysis to uncover similarities and differences (n=22). After abstract screening of all review papers, n=10 review papers were discarded because they didn't fit in inclusion criteria of this paper. Thus, n=12 review papers were selected for full-text assessment to determine eligibility. Table 3.1 presents the review papers that fit inclusion criteria defined for this research.

Table 3.1: Review Articles

Title	Authors (year)	Review focus	Conclusions
Survey on Data Analysis in Social Media: A Practical Application Aspect [12]	Hou, Qixuan; Han, Meng; Cai, Zhipeng (2020)	This research conduct a systematic comparison of existing applications, categorizing them by analysis techniques and areas of impact, with the aim of offering a comprehensive and detailed survey of social media-based applications. It explores how social media applications impact healthcare, disaster management, and business operations.	Research outlines a comprehensive four-stage pipeline for building social media-based applications. Those four stages are data collection, data storage, data analysis and data visualization. Research emphasize key analysis techniques such as topic analysis, time series analysis, sentiment analysis and network analysis and their applications. It also highlights the importance of addressing privacy concerns, the impact of 5G technology, and suggests future research directions to enhance and innovate within the field.

Title	Authors (year)	Review focus	Conclusions
A systematic review of big data and digital technologies security leadership outcomes effectiveness during natural disasters [2]	Adegoke, Damilola (2023)	This review primarily aims to research the current viewpoints, discoveries, and stances presented in existing literature concerning the effectiveness of leadership decision-making outcomes in crisis or disaster situations, specifically focusing on big data and digital technology security.	The review shows that social media, particularly Twitter and Facebook, play a dominant role in security leadership and natural disaster management compared to other digital technologies. The studies highlight five main themes: big data in crisis decision-making, crisis communication, disaster preparedness, disaster recovery, and social media as a public space during disasters. However, the findings also point out a gap in the literature, emphasizing the need to explore instances where social media has resulted in ineffective outcomes.
How Advanced Technological Approaches Are Reshaping Sustainable Social Media Crisis Management and Communication: A Systematic Review [13]	Bukar, Umar Ali; Sidi, Fatimah; Jabar, Marzanah A.; Nor, Rozi Nor Haizan; Abdullah, Salfarina; Ishak, Iskandar; Alabadla, Mustafa; Alkhalifah, Ali (2022)	The focus lies in assessing how machine learning (ML), social network analysis (SNA), and associated solutions are transforming sustainable crisis management and decision-making processes, drawing insights from existing literature.	Social media plays a crucial role in crisis informatics by facilitating the collection and dissemination of diverse information for crisis management. The study highlights that advanced technologies, including ML techniques and network analysis tools, significantly enhance sustainable crisis response efforts, although challenges like big data handling and cross-platform support remain.

Title	Authors (year)	Review focus	Conclusions
Earthquake Reconnaissance Data Sources, a Literature Review [14]	Contreras, Diana; Wilkinson, Sean; James, Philip (2021)	The objective of this paper is to pinpoint cutting-edge data reservoirs suitable for constructing damage assessments and offer advice on optimizing data gathering processes for enhanced efficiency.	This review has highlighted the most commonly used data sources in earthquake reconnaissance over the past years and the major seismic events within that timeframe. Traditional methods, like fieldwork and remote sensing (RS), have evolved, incorporating new technologies for greater accuracy and efficiency. Crowdsourcing and social media platforms, are now increasingly significant data sources, providing first-hand insights and complementing official data collection efforts. Additionally, researchers must combine various data sources to ensure comprehensive and reliable assessments of seismic events, as no single method suffices on its own.

Title	Authors (year)	Review focus	Conclusions
Using Mobile Phone Data for Emergency Management: a Systematic Literature Review [15]	Wang, Yanxin; Li, Jian; Zhao, Xi; Feng, Gengzhong; Luo, Xin (Robert) (2020)	This systematic literature review examines 65 studies that explore the utilization of mobile phone data for emergency management. It puts forward a framework aimed at consolidating the dispersed knowledge from these existing studies.	Emergencies impact society by causing economic losses and human casualties, challenging traditional management methods that rely on limited data sources. Despite its contributions, this study has several limitations. First, like many literature reviews, it faces constraints related to keyword selection, though efforts were made to mitigate this by drawing from prior research in emergency management (EM) and mobile phone data. Second, the proposed framework is based solely on the reviewed articles, so it may overlook unmentioned relationships. Additionally, while comprehensive, this review does not cover emerging methods or provide a detailed exploration of data processing and analysis techniques.

Title	Authors (year)	Review focus	Conclusions
Emergency Decision Making: A Literature Review and Future Directions [16]	Su, Wenxin; Chen, Linyan; Gao, Xin (2022)	This study utilizes a literature review to both summarize the advancements in research and pinpoint future directions in Emergency Decision Making (EDM). It holds both theoretical and practical significance, presenting a theoretical framework for EDM rooted in stakeholder theory.	The study of EDM is vital for managing crises and achieving sustainable development goals, but its diverse findings make it challenging to understand the current research. This literature review summarizes progress and establishes a stakeholder-based theoretical framework, suggesting future research in four areas: social media analysis in emergencies, improved computer-aided tools, the influence of decision-makers' characteristics, and better coordination of stakeholders in emergency projects.
Edge Technologies for Disaster Management: A Survey of Social Media and Artificial Intelligence Integration [17]	Aboualola, Mohamed; Abualsaud, Khalid; Khattab, Tamer; Zorba, Nizar; Hassanein, Hossam S. (2023)	The survey provides a thorough examination of prior literature pertaining to the subject matter, organized into four distinct phases. It outlines a coherent approach for managing emergency situations, incorporating various cutting-edge technologies such as sensing, IoT technologies, big data, AI, and SM analytics within each phase.	The integration of advanced technologies like IoT, AI, and big data analytics can significantly reduce casualties and infrastructure damage during crises. This survey reviews recent research on SM and AI-based emergency management systems, highlighting current disaster management technologies and their suitability for crisis scenarios. While many benefits are identified, the survey also points out areas that still require further improvement.

Title	Authors (year)	Review focus	Conclusions
Toward an Integrated Disaster Management Approach: How Artificial Intelligence Can Boost Disaster Management [18]	Abid, Sheikh Kamran; Sulaiman, Noralfishah; Chan, Shiau Wei; Nazir, Umber; Abid, Muhammad; Han, Heesup; Ariza-Montes, Antonio; Vega-Munoz, Alejandro (2021)	This paper emphasizes the critical role that AI can play in enhancing disaster management across its various phases: mitigation, preparedness, response, and recovery.	This research explores how AI enhances disaster management by reviewing various AI applications across disaster phases. It emphasizes the growing role of geospatial technology, GIS, and RS in understanding and mitigating disasters. The study concludes that as AI technology and multispectral datasets evolve, they will become even more effective in reducing disaster impacts, though success also depends on robust data management and analytical capabilities.
A systematic review of natural language processing applications for hydrometeorological hazards assessment [19]	Tounsi, Achraf; Temimi, Marouane (2023)	This paper examines reviewed studies, including journal articles and conference proceedings, that utilized NLP to analyze extreme weather events, with a particular emphasis on heavy rainfall incidents.	The findings indicate that NLP is still not used enough in the study of extreme weather events, though it has the potential to significantly enhance the value of data from social media, newspapers, and other sources for weather assessment. However, for NLP to be effectively employed, challenges such as data inadequacy, accessibility issues, non representative methods, and the need for advanced computing skills must be addressed.

Title	**Authors (year)**	**Review focus**	**Conclusions**
Social media for intelligent public information and warning in disasters: An interdisciplinary review [20]	Zhang, Cheng; Fan, Chao; Yao, Wenlin; Hu, Xia; Mostafavi, Ali (2019)	The author visualize an intelligent public information and warning system for disasters that leverages social media, with three primary functions: efficiently and effectively gathering situational awareness data during disasters, facilitating self-organized peer-to-peer assistance, and allowing disaster management agencies to receive input directly from the public.	This paper outlines a vision for intelligent public information and warning systems during disasters, highlighting three key functions. It reviews existing studies to identify challenges in using social media for disaster communication and proposes a research roadmap to overcome these obstacles, ultimately aiming to enhance societal resilience through improved social media use.

Title	Authors (year)	Review focus	Conclusions
Managing natural disasters: An analysis of technological advancements, opportunities and challenges [21]	Krichen, Moez; Abdalzaher, Mohamed S.; Elwekeil, Mohamed; Fouda, Mostafa M. (2023)	The authors provide an insight into how using technologies such as RS, IoT, SM can be used to predict, react and recover more efficiently within disaster management.	The integration of technologies like RS, satellite imaging, IoT, smartphones, and SM can significantly enhance natural disaster management by providing real-time data, visualizations, and analyses, which improve decision-making and emergency response. However, challenges such as implementation costs, data privacy concerns, and the need for skilled personnel must be addressed, requiring collaboration between organizations and governments, with future efforts focusing on making technologies more affordable, developing standardized protocols, and promoting cross-sector cooperation.

Title	Authors (year)	Review focus	Conclusions
How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions [22]	Sun, Alexander Y and Scanlon, Bridget R (2019)	This survey aims to explore the potential and advantages of data-driven research in Early Warning Management (EWM), summarize key concepts and methodologies in Big Data and Machine Learning, systematically review current applications, and address significant challenges and issues while proposing future research directions.	The review of over 1000 papers highlights the transformative impact of Big Data on environmental and water management (EWM) research, emphasizing the need for automated data wrangling, accessible data cleansing, and the application of deep learning (DL) techniques for advanced analytics. However, challenges such as data cleaning, lack of labeled datasets, mismatched data speeds, high costs, and gaps in data governance must be addressed to fully leverage these technologies.

From this analysis, it can be concluded that no previously published review paper fully addresses the research question examined in this paper.

4. NATURAL DISASTERS

Various natural hazards can be classified into three primary categories: meteorological, hydrological, and geological hazards. Meteorological events cover phenomena such as tornadoes, hurricanes, thunderstorms, winter storms (including ice storms), and summer storms (including wildfires). Hydrological events include different types of floods (fluvial, pluvial, and coastal), storm surges, and tsunamis. Geological hazards comprise earthquakes, volcanic eruptions, and mass movements such as landslides, mudflows, and avalanches [23]. The most common natural disasters in the world include floods, earthquakes, cyclones, hurricanes and typhoons, droughts, fires, and landslides. To address the first research question, we screened the papers included in this analysis and identified the type of disaster that each paper addresses.

Figure 4.1 shows a comparison of the average occurrence of natural disasters from 2003 to 2022 with the average occurrence of disasters in 2023. Floods are most current natural disaster judging by this picture. A review of the literature reveals that floods are the most frequently mentioned natural disaster, as illustrated in Figure 4.2. The 'Storm' category encompasses hurricanes, typhoons, winter storms, and tsunamis. Other categories depicted in Figure 4.1 which are absent from Figure 4.2 were excluded because they were not the focus of our research.

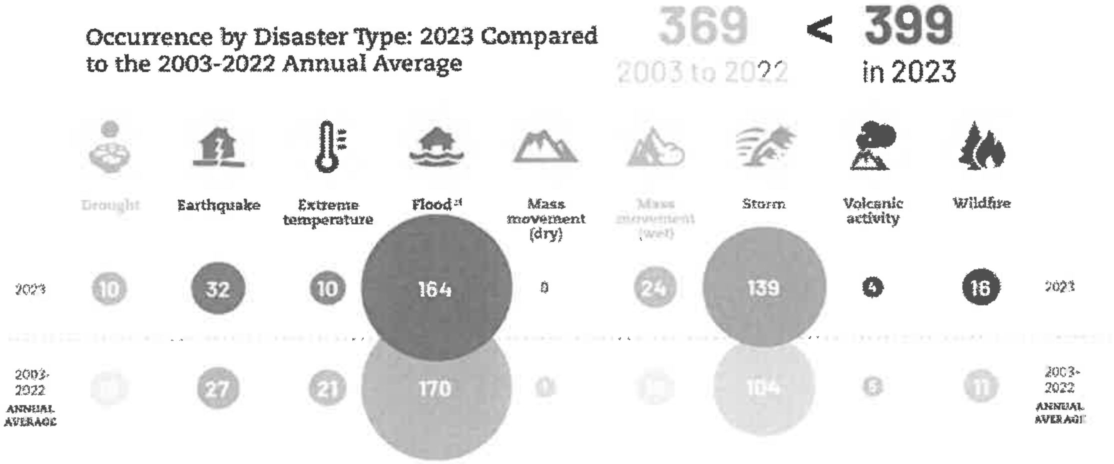

Figure 4.1: Natural disasters in 2023 [24]

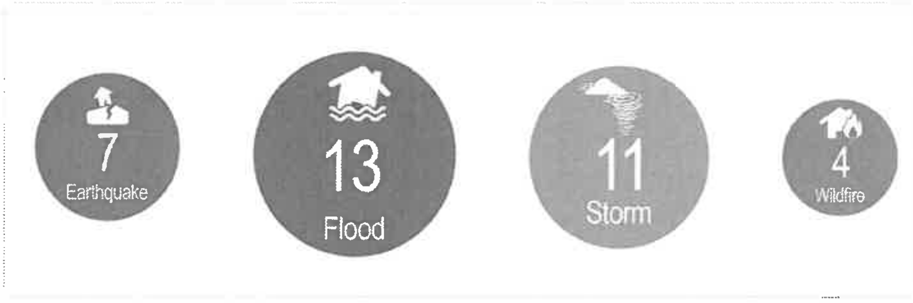

Figure 4.2: Natural disasters mentioned in literature

When comparing Figure 4.1 and Figure 4.2 there is overlapping between the frequency of natural disasters that happened during years and how often they are discussed in literature. For instance flood is most frequent natural hazard in world according to Figure 4.1 and it is also most mentioned in literature included in this research. Following floods, storms are the next most frequently mentioned hazard in the literature, which aligns with their high occurrence rate globally. After storms, earthquakes come next in both real-world frequency and literature mentions. Finally, wildfires are the least discussed in the literature, which corresponds to their lower frequency compared to other hazards. This order of mention in the literature closely follows the actual occurrence of these natural disasters, indicating a certain level of alignment between real-world data and academic focus.

5. DATA SOURCES

In this section, different sources of textual crisis data will be introduced, focusing on their significance in disaster research. Volunteered Geographic Information (VGI) platforms, social media networks, official reports, and newswire services each play a critical role in providing timely and valuable data during crises. VGI platforms offer georeferenced data contributed by individuals, while social media channels provide real-time user-generated content, such as text, photos, and videos, that can be quickly analyzed for insights. Official reports from government agencies and international organizations deliver authoritative information but often lag in real-time updates. Newswire services, including online news portals, contribute structured, formal data, complementing the immediacy of social media with in-depth analysis and verified facts.

5.1. Volunteered Geographic Information

Volunteered Geographic Information (VGI) are become known in scientific applications about natural hazards [25] in past decade. VGI is actually georeferenced data provided by individuals thought collaborative crowdsourcing activity. With specialized tools they can be collected, analyzed, shared for some purposes such as scientific researches about natural disasters. VGI data can be posted in different forms, such as text, photos or videos. It is a cost-effective alternative to traditional data collection methods, offering near real-time information and location data that can be used to identify disaster risk areas and generate hazard maps [26]. That data can give very useful information from the ground in cases when satellite or airborne imagery can't. Today, there are online platforms that are developed for collecting data for purposes of citizens science. Those platforms carry data for limited period and for certain events. For example online platform "Did you feel it?" (DYFI) [27] is developed by United States Geological Survey (USGS). Purpose of this platform is to collect and analyze firsthand information of earthquake experiences from people who felt the tremors. In Europe there are several online platforms for monitoring and providing real-time information about different kind of phenomena based on VGI. European-Mediterranean Seismological Centre (EMSC) - LastQuake [28] that provides real-time earthquake information and collects user reports. Copernicus Emergency Management Service (EMS) [29] where users can contribute observations and reports on disasters such as floods and fires. The platform integrates VGI for validation and mapping purposes. Flood-CitiSense [30] enables users to report flooding events, participate in data collection, and access real-time flood information. For instance, the authors [26] mention an online document tool called Shimo [31] that is used to collect data for this paper research. This tool allows users to compile and share information from platforms like Sina Weibo [32] and WeChat [33] in an open-access document accessible to all.

5.1.1. Social Media

When there is talk about natural disasters or any kind of emergency in our close environment, we instantly go on some Social Media (SM) to see what people write about it. SM in emergency is very popular these days and it can give a lot of useful information in real time.For example, in [34] authors noted that first tweet about California Morgan Hill earthquake was posted 19 s after the earthquake, while official data was posted 22 s after earthquake.

With every day there are more and more individuals connected on some social network.Web page DOMO "Data never sleeps" [35] tracks information about number of people in the world that use social networks, how many tweets, FB posts etc. are generated in every minute. Figure 5.1 presents the usage of social networks in 2021 as reported on the DOMO page. According to their reports, in 2021, Facebook users posted 240,000 photos every minute. Twitter saw 575,000 tweets per minute. The same report show that the number of internet users increased from 3.4 billion to 5.2 billion, from 2016 to 2021.

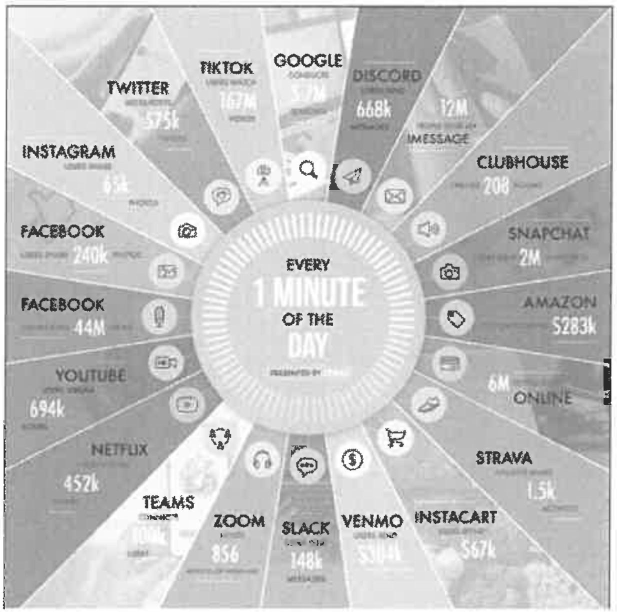

Figure 5.1: Usage of social networks in 2021 [35]

There are four key areas of SM data that are objects of researchers: disaster response, public sentiment analysis, human activity studies, and disease prevention and control [36]. Additionally, many applications analyze social media to offer valuable insights in this field.

In the following subsections, the social networks referenced in the analyzed papers will be thoroughly described.

Twitter

When screening the literature for the purpose of this review, it is noticed that most papers use the social network Twitter. Twitter [4] is the most commonly used network in social media data analysis due to its publicly accessible data, robust API, and real-time nature, which make data collection and analysis straightforward and efficient. The platform's short, text-based posts, limited on 280 characters, facilitate natural language processing, while the diversity of topics discussed allows for a wide range of research subjects. Character limitation on posts and messages enables the extraction of big data that is easier to process, analyze, and evaluate [37]. Additionally, the rich metadata accompanying each tweet, including timestamps, geolocation, and user interactions, enhances the depth of analysis. Twitter's network structure, characterized by followers, retweets, mentions, and hashtags, provides a valuable framework for studying information dissemination and social networks. Furthermore, the availability of historical data supports longitudinal studies, enabling researchers to track trends and shifts in public opinion over time. These attributes collectively make Twitter a highly attractive and practical choice for researchers in social media data analysis. Twitter introduced an "academic research product track" in January 2021 that gived access to 10 million monthly records from its historical data repository [38], [6]. Today, Twitter's free tier is extremely limited, offering access to only 1,500 tweets per month. In 2023, Twitter was renamed to 'X' as part of the company's rebranding efforts to create a platform that offers a wide range of digital services beyond just social networking. After rebranding, Twitter announced major changes to its API, including the introduction of paid tiers and the discontinuation of free access which actually led to major changes in the course of research by various scientists. If Twitter doesn't change its policy, maybe that can be end of an era of research this kind of data [39].

Sina Weibo

Sina Weibo [32], or just Weibo, is one of China's most popular social media platforms. It can be compared to a hybrid of Twitter and Facebook and serves as a microblogging site where users can post updates, share content, and engage with others. This social network was launched by Sina Corporation in August 2009. Users can post text, images, videos, and links, similar to Twitter. Posts can be up to 2,000 characters long, allowing for more detailed content compared to Twitter. Weibo is a powerful tool for information dissemination and public discourse in China. It has played a crucial role in spreading news, mobilizing public opinion, and facilitating discussions on social issues.

Facebook

Facebook [5] is one of the largest and most widely used social networks globally. Unlike Twitter and Weibo, which are primarily text and microblogging platforms, Facebook offers a comprehensive social networking experience with a variety of features that include status updates, photo and video sharing, events, and private messaging. Its extensive and diverse user base makes it a rich source of data for social media research. Researchers value Facebook for its detailed user profiles and the complex network of friendships and interactions, which provide insights into social connections and behaviors. The platform's API, though more restricted than Twitter's, still supports significant data extraction for research purposes. Facebook's ability to facilitate long-form content and multimedia sharing, combined with its robust advertising and engagement tools, makes it a powerful platform for studying social interactions, marketing strategies, and information dissemination. Additionally, the historical data available on Facebook allows researchers to perform in-depth longitudinal studies, tracking changes in social dynamics and public sentiment over time.

Crimson Hexagon

Crimson Hexagon [40] is a SM analytics platform that leverages advanced ML and AI to provide deep insights into social media data. Unlike platforms such as Twitter and Weibo, which are primarily social networks, Crimson Hexagon is a tool used by researchers and marketers to analyze social media content across various platforms, including Twitter, Facebook, Instagram, and more. It excels in sentiment analysis, trend identification, and audience segmentation, allowing users to understand public opinion and consumer behavior on a large scale. The platform's ability to analyze historical data and real-time streams enables comprehensive longitudinal studies and immediate insights into current events. Crimson Hexagon's extensive database and robust analytical tools make it an invaluable resource for social media research, offering capabilities that go beyond the native analytics of individual social networks. Some authors use Crimson Hexagon as a tool for accessing data from various social media platforms.

5.2. Official reports

In the scope of this study official reports refers to records where sources are government agencies [41], meteorological institutes, international organizations and similar authorities. Civil defense organizations have valuable databases that contain information about natural disasters [42]. Official records can be used for creating reports and analysis on the frequency and severity of natural disasters. They are important in damage estimation after natural disaster but also for creating strategies for prevention of risks. Although official data are usually collected and verified by experts, and they adhere to strictly defined standards and methodologies their flaw is that collection, verification and distribution may take time, which means that it may not be available in real time. Also, this kind of data can be unavailable for public because of some political, security or legal reasons. A combination of official data and data from social networks can provide the most complete picture of natural disasters. Official data offers a reliable basis, while social networks enable speed and local insights. It is ideal to use both sources with appropriate verification and analysis to ensure accuracy and comprehensiveness of information.

5.3. Newswire

When referring to traditional media services as data source, in the context of disaster management there can be mentioned TV media, newspapers, online portals with daily news. Since this paper is related to textual data than we can refer to services that are presented with writing. Online platforms that offer daily news range from global giants to local-focused websites, each providing different scopes and styles of reporting are available in each country in the world. From those portals we can also collect data that can be taken as research data. Data for news are usually written in structured form for publishing in newspapers or online platforms.

Paper [43] propose method to compare characteristics of newswire and SM. A comparison of the word clouds used for posts about Nepal earthquake shows that while newswires focus on formal, objective language related to the Nepal earthquake, using terms like "government," "reconstruction," and "political," tweets highlight the human aspect, offering complementary perspectives that together provide a more comprehensive understanding of the disaster, with the immediacy of tweets being valuable for first responders in search and rescue operations. Also authors came to conclusion that most news about disaster appear first on Twitter and then on newswire.

5.4. Social media platform selection for crisis management

The diverse data sources outlined in this section—Twitter, Sina Weibo, Facebook, Crimson Hexagon, official reports, and newswire services—each offer unique advantages and challenges for data collection and analysis, particularly in the context of social media research and disaster management. Official reports remain crucial for validated and reliable data, although they may be delayed and limited in scope. Newswire services contribute structured, professional data that complement both social media and official sources. VGI, through crowdsourced geographic data, allows for real-time, location-based insights during disasters. Moving forward, effective data preprocessing, such as cleaning and NLP, will be essential to integrate and analyze these diverse data streams. Combining real-time social media and VGI data with verified official records offers the most comprehensive approach to disaster management and response planning. During disaster people posts great amount of text, pictures and videos that can actually be useful for several reasons:

- SM enables quick and widespread distribution of information, which is crucial in emergency situations. People can learn about dangers, evacuation routes, and shelter locations in real-time.
- People can use SM to coordinate help and support. For example, volunteers can be organized, donations collected, or resources such as water, food, and shelter shared.
- SM posts can draw the attention of the wider public and international organizations to affected areas, leading to faster and more effective mobilization of humanitarian aid.
- Through user posts, organizations and authorities can gather data about the situation on the ground, which can help in better understanding the situation and planning relief efforts.
- Using SM to share accurate and verified information can help combat the spread of panic and misinformation that can make worse the situation during natural disasters.

Due to the extensive use of social media, numerous scientific papers have been published on data extraction in emergency situations. Figure 5.2 shows the usage of social media in scientific papers expressed in percentage. It can be observed that the most used social media in papers is Twitter, while the Official reports and Crimson Hexagon were only used in 5% of publications.

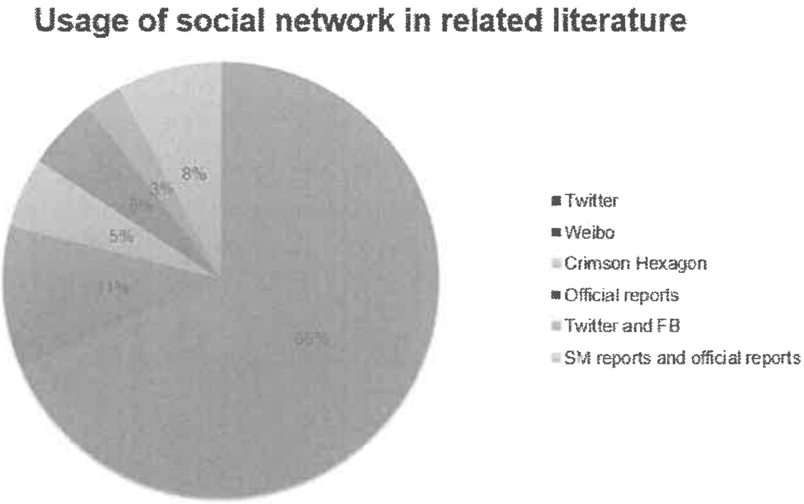

Figure 5.2: Usage of social media in papers

6. MAIN ANALYTIC TECHNIQUES

Most of the reviewed literature highlights common steps involved in processing data from social media. First, data is extracted from specific web pages using methods such as web scraping tools or APIs. The raw data then undergoes preprocessing, which typically includes cleaning, tokenization, and filtering to remove irrelevant or noisy information. After preprocessing, the data is further processed to yield structured, useful information, often involving the application of algorithms for tasks such as clustering, sentiment analysis, or classification.

In many cases, the temporal nature of social media data plays a crucial role in analysis. Time series analysis becomes important when dealing with sequential data, where patterns and trends over time provide deeper insights.

This section describes various methods for data extraction, along with commonly used techniques for data preparation, such as text normalization and feature selection. Additionally, it explores algorithms frequently applied to process refined data, highlighting both their benefits and the challenges encountered when analyzing large volumes of unstructured social media data. The combination of different data types and sources, as well as the algorithms used for processing them, will also be discussed. Time series analysis and related measures, which play a significant role in social media text analysis, are presented as a key component of the analytical techniques explored in this research.

Structure of this section is demonstrated in Figure 6.1.

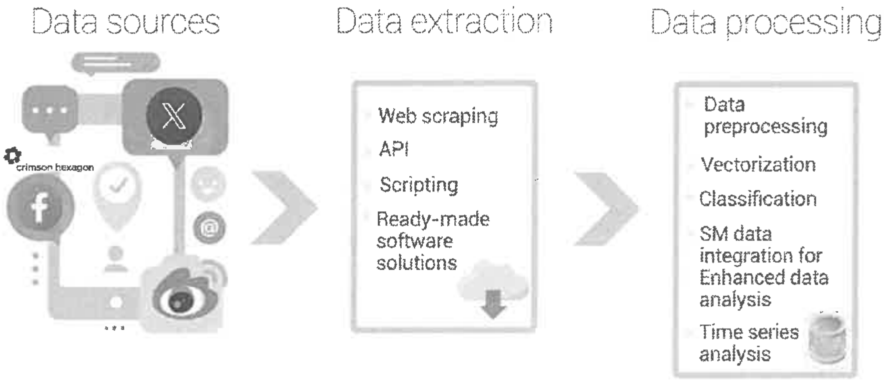

Figure 6.1: Structure of main analytics techniques

6.1. Data extraction

Web scraping is automated technique that allows to extract large amount of unstructured data from some web page [44]. Most popular way of using this technique is thought online services, specific API's. Another way is to create your own script for downloading data. Of course, it is necessary to have permission to get data from web page. Web pages like Twitter or Facebook have API's that gives structured data once when we access them. But many pages gives unstructured data when other advanced technologies are used for processing.

The crawler and the scraper are main parts of every web scraping. **Crawler** refers to software capable of systematical browsing of web pages. It is algorithm that search through web following links to find specific data as needed. the process begins with a small set of links, and the crawler identifies additional links on those pages, adding them to a queue, known as the crawl frontier, for further exploration. To function efficiently, a good crawler should be able to detect circular references and slight variations of the same page [45]. The second mentioned, **scraper**, helps to extract data from certain web page. Web scraper follows given URL and than it loads all HTML elements from specific site, also known as HTML parsing. Web scrapers use parsers to extract useful information from collected data. Most web scraping libraries include basic HTML parsing, and specialized parsers exist for formats like PDF, CSV, QR code, and JSON. Real web browsers, such as Firefox and Chrome, have built-in parsers that web scraping tools can also leverage [45]. It is essential to specify what data we want to extract so the scraper can do his job more quickly and more efficiently.

Beautiful Soup is a popular web scraping tool for parsing HTML, which involves extracting data from HTML and XML documents. It's especially useful for retrieving structured information from static HTML. One of its main advantages is its ability to efficiently navigate HTML structures and extract the desired content using various filtering and parsing methods [46].

Geo Tweets Downloader (GTD) is software that uses Twitter API's for downloading twitter georeferenced posts and saves them in PostgreSQL in real time and it's chosen in paper [47] for data extraction.

In the papers, data extraction and analysis were conducted using ready-made software solutions such as Octoparse [48] and the Social Media Analytics Framework (SMAF) [49]. Both Octoparse and SMAF are designed to extract data. Octoparse focuses on web scraping and automation, allowing users to extract data from websites without coding. SMAF, on the other hand, is specifically tailored for extracting data from various social media platforms. The use of such software solutions can expedite the processes of data extraction and analysis, particularly in situations where it is necessary to gather large amounts of information from websites or social media. Of course, manual extraction of data can also be an option but it is recommended only for small amount of data.

6.2. Text processing

Natural Language Processing (NLP) is part of AI that helps computer to understand human language and at the end, to respond, react on some meaningful way to both sides. It combine ML algorithms, deep learning models and neural networks with computational linguistic. NLP consists of various process including sentiment analysis, speech recognition, language translations etc. Syntactic techniques are focused on sentence structure, the grammatical structure of sentences to understand how words are organized and how they relate to each other. Parsing is related to understanding the meaning of the words. Semantic analysis is often used with sentiment analysis. It's role is to interpret meaning of words or phrases. To understand language machine must understand the context of the text and it often depends on other words that are closed to each other and broader conversation. At the end machine respond in way of translation or some other task based on their textual understanding. There are three main phases in text processing that were majority of papers included in this study follow:

- data preprocessing
- vectorization
- classification

All of these phases will be described in this section and algorithms and methods that are connected to this phase.

Analyzing social media data during disasters is tough because it's so diverse. Social platforms churn out a lot of information in emergencies, making it hard to find what's important. Dealing with this flood of data needs advanced tools. Also, social media content comes in many forms, like text, images, and videos, which makes it tricky for traditional analysis methods. For example, text often has slang and mistakes, making it tough for computers to understand. Plus, there's a lot of noise in social media data, so filtering out irrelevant stuff and finding credible sources is crucial. Social media posts often look like casual speech, so getting useful info from them means using different techniques. Overall, analyzing social media during disasters is a big challenge that needs smart strategies to tackle. Extracting meaningful information from such unstructured textual content necessitates the consideration and application of diverse preprocessing techniques.

6.2.1. Text preparation-preprocessing

After data extraction, next step is data preprocessing because SM data is full of noisy and unnecessary data. In reviewed paper there are mentioned some of main NLP preprocessing methods:

- **Keyword extraction**- involves automatically identifying the key terms that best represent the content of a document. Various terms are used to define these key elements, such as key phrases, key segments, key terms, or simply keywords. Despite the different terminology, their purpose remains the same: to identify the main topics within a document [50].
- **Parsing**- refers to the process of resolving a sentence, word, or paragraph into its main components or logical parts according to a set of predefined rules or structures [51].
- **Regular expression (RE)**-Regular expressions, besides being a key concept in formal language theory, are extensively used in computer science to define search patterns [52].
- **Tokenization**- one of the key components in NLP. That means that text need to be splitted in smaller parts-tokens, like words or sentences [53]. All not necessary details need to be removed from text, like punctuation or special letter, so that machine can understand human language.
- **Stemming**- process of generating different forms of a root or base word. Simply put, it reduces a word to its stem or root form. This technique is used to simplify search queries and standardize sentences, making them easier to interpret [54].
- **Lematization**- The process involves grouping the inflected forms of a word so they can be recognized as a single element, known as the word's lemma or vocabulary form. Unlike stemming, which focuses on reducing words to their root, lemmatization adds meaning by linking words with similar meanings to a common lemma. It uses an algorithm to identify the lemma based on the word's intended meaning rather than just its root form [54].

[55] is based on extracting keywords related to hazard that is researched, and to understand social impact of hydrometorological events. The paper [56] focuses on extracting detailed road condition information from social media text. It introduces a method that combines NLP with a semantic knowledge base, overcoming challenges like short and spoken-like social media text. The method uses rule templates to extract hazard damage information, matching candidate words against a knowledge base enriched with third-party resources. Experimental results confirm its effectiveness, particularly in extracting information during typhoon disasters.

Also, Natural Language Toolkit (NLTK) is a popular and useful tool for data preprocessing.The authors of the paper [23] focus on eliminating irrelevant information and reducing redundancy, both of which are crucial for most natural language processing tasks. A Python script was employed, utilizing regular expressions (RE) and the NLTK libraries. Stemming and lemmatization were applied to derive the root or base form of words, facilitating subsequent feature extraction and processing. Stemming was followed by feature extraction using the bag-of-words method, which involves building a vocabulary of words in the dataset and computing the occurrence of each word in each data instance, such as a single tweet in this case. Table 6.1 summarizes various preprocessing techniques used in data analysis in various studies from the literature, such as data cleaning, keyword extraction, normalization, regular expressions (RE) and tokenization.

Reference	Data cleaning	Keyword extraction	Normalization (lemm. OR/AND stemm.)	RE	Tokenization
[6]		x			
[23]			x	x	
[55]		x			
[56]	x	x			
[57]		x			
[58]	x	x			
[59]	x	x	x	x	
[60]	x	x	x		
[61]	x		x		x
[62]	x				
[63]	x	x	x		x
[64]		x			x
[65]	x		x		x

Table 6.1: Techniques used for preprocessing data

6.2.2. Vectorization

Text analysis involves turning unorganized documents into structured data. In machine learning, especially supervised learning, this means creating feature vectors, a process called vectorization. In natural language processing, there are two main ways to do this:
Mapping text to string-based vectors, where words or sentences are directly represented.
Mapping text to number-based vectors, where words are represented indirectly. The string-based method is rarely used due to the difficulty of comparing words as strings, but it is easier for humans to understand and useful for spotting errors. The number-based method, including techniques like word embeddings, is more common and involves converting words into numbers [57].

- **Bag of words (BoW)**- is a simple text representation method commonly used in NLP. It represents text as an unordered collection of words, where the focus is on the presence or frequency of words rather than their order or context. Each word is treated as an independent feature, allowing for basic text analysis, such as word counting [58].
- **Embedding**- Vector semantics involves representing words as points within a multidimensional space, where their positions are determined by the distributions of neighboring word embeddings. These vectors, used to represent words, are referred to as embeddings. The term "embedding" comes from its mathematical meaning, signifying a mapping from one space or structure to another [59].
- **Term Frequency–Inverse Document Frequency (TF-IDF)** - a statistical technique used to assign weights to words in a text. It involves two main components. The first is term frequency, which reflects how often a word occurs in a document, adjusted to account for the document's length. This prevents longer documents from skewing the word counts. The second component is inverse document frequency, which assesses how significant a word is across multiple documents. While term frequency treats all words equally, inverse document frequency reduces the weight of common words that may frequently appear but contribute little to the overall meaning of the text.

A popular neural language model is Word2Vec, which is trained in an unsupervised, generative manner to convert text documents into feature vectors. However, like TF-IDF, Word2Vec operates at the phrase level and is typically trained on well-organized text, making it less suitable for the informal nature of Twitter content. Tweet2Vec addresses these limitations by using a character-based model, which is trained self-supervisedly through a hashtag prediction task. Moreover, unlike the shallow neural network in Word2Vec, Tweet2Vec employs a more advanced approach [60]. So, in this study [60], three text processing methods are employed for transforming tweets into numerical vectors. Keyword matching identifies tweets related to flooding using specific keywords. The TF-IDF approach converts tweets into vectors based on term frequency, with preprocessing steps such as removing user references and expanding hashtags. The Tweet2Vec method uses a bidirectional LSTM to create dense embeddings, handling Twitter-specific language artifacts through character-based processing and tailored preprocessing steps. These methods enable effective analysis of tweets to estimate flooded areas.

Topic modeling is a type of statistical modeling used in NLP to discover abstract topics within a collection of documents. It helps in understanding the themes or subjects that are prevalent in a large set of text data. Two key points are related to topic modeling: topic and document-term matrix. In topic modeling, each document is represented as a mixture of topics, and each topic is represented as a mixture of words. Document-Term Matrix is a matrix where each row represents a document, and each column represents a term (word). The values in the matrix typically indicate the frequency of the term in the document. Most common algorithms for topic modeling are:

- Latent Dirichlet Allocation (LDA) - It is introduced in 2003 by [61]. LDA is an unsupervised Bayesian algorithm that assumes latent topics within data, with each topic being a probability distribution over words. However, it requires a predefined number of topics, where fewer topics result in broader categories, and more topics can lead to overlapping themes [62]. Assumes that documents are generated from a mixture of topics, and topics are generated from a mixture of words. Uses a probabilistic approach to find the mixture of topics in each document and the mixture of words in each topic [63].,

- Non-Negative Matrix Factorization (NMF) - A matrix factorization technique that decomposes the document-term matrix into two lower-dimensional matrices. Ensures that the resulting matrices have non-negative values, which makes it easier to interpret the topics.,

- Latent Semantic Analysis (LSA). - Latent Semantic Analysis (LSA), also referred to as Latent Semantic Indexing, is a statistical approach for uncovering and representing the hidden semantic structure within a set of documents. Unlike Latent Dirichlet Allocation (LDA), which is a generative probabilistic model, LSA is based on matrix factorization. Its primary goal is to capture patterns of word co-occurrence across documents. LSA operates under the assumption that words appearing together in similar contexts tend to share similar meanings [64].

To derive valuable insights and common indices that characterize the spatiotemporal and socioeconomic patterns of social media activity during meteorological disasters, authors of paper [65] employed comprehensive research framework. This framework integrates three key components: data preparation, theme mining using the Latent Dirichlet Allocation (LDA) model, and subsequent statistical analysis. Data was taken from Sina Weibo social network and LDA model was used to extract themes from that posts. Subsequently, the development and spatial distribution of these themes were explored and interpreted. Statistical analysis was then utilized to discern thematic variations between regions that were severely impacted and those that were generally affected. Finally, an analysis of the socioeconomic disparities in the themes was conducted in conjunction with socioeconomic data. There were identified four themes and each theme had keywords connected to them. Also, these authors made map of spacial distribution of each theme. Paper [64] emphasizes that preprocess step was crucial for further data analysis. Utilizing Python packages like Pandas and Numpy, the dataset, initially in Tab-separated values (TSV) format, went through steps such as removal of duplicates based on tweet IDs and splitting based on label text. This process allowed for a focused analysis on specific events and categories, providing a granular understanding of the types of damage, impacts, and humanitarian responses. After preprocessing the data three distinct topic modeling techniques were employed: Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF). These techniques aimed to reveal the underlying topics and themes present in the tweets concerning the disasters. By organizing and refining the dataset through preprocessing steps, such as duplicate removal and categorization, the analysis focused on extracting meaningful insights and understanding the key aspects and narratives within the tweets.

6.2.3. Social media posts classification

After vectorization, machine learning models are used for various tasks such as classification (e.g., text categorization), clustering, entity recognition, or regression. In this step, the results of the models are analyzed, and decisions or estimations are made based on the numerical representations of the text.

Classification techniques - A critical phase in the text classification process is selecting the most appropriate classifier. One of the simplest classification algorithms is logistic regression (LR), which has been extensively utilized across various data mining domains. The Naïve Bayes Classifier (NBC) is particularly popular due to its computational efficiency and minimal memory requirements. Non-parametric methods have also been explored for classification tasks, such as the k-nearest neighbor (KNN) algorithm. Support Vector Machines (SVM), another widely adopted technique, employs a discriminative approach to document categorization. SVM has applications across numerous fields, including bioinformatics, image and video analysis, human activity recognition, and security. Often, researchers use SVM as a baseline to evaluate the performance of their models, highlighting their contributions and innovations. Tree-based classifiers like decision trees and random forests have also been extensively studied for document categorization, with each tree-based method discussed in its respective sub-section [66]. Additionally, recent advancements have seen the rise of graphical models

for classification, such as conditional random fields (CRFs), which are frequently employed in tasks like document summarization and automatic keyword extraction [67].
The authors of the paper [68] conducted research on the 2014 South Napa earthquake. Data was extracted from Twitter posts and preprocessed through several steps: tokenization, case transformation, removal of stop words, elimination of short words, and stemming. After preprocessing, the data was classified using a SVM to identify damage-related messages. The accuracy of the classification was evaluated using a confusion matrix. The results indicated that the SVM classifier effectively identified damage-related messages. Based on the results, the authors created a damage map. NMF topic modelling was used in paper [69]. Tweets were devided in two categories: Tweets that are related to natural disasters and Tweets that are not related to natural disasters.

In this paper, NMF is used to break down a matrix of tweet data into simpler parts to find hidden topics in crisis-related tweets. This helps uncover what people are talking about and how these topics change over time and across locations, making the information easier to understand. while the spatial distribution of tweets associated with specific topics like "flood" was mapped to visualize their occurrence and impact over time.
Sentiment analysis, also known as opinion mining, is a NLP technique used to determine the emotional positive or negative tone behind a body of text. This analysis helps identify the attitude expressed in a piece of text, which can be classified into positive, negative, or neutral sentiments. Two mayor approaches of sentiment analysis are Lexicon-based approach and Learning-based approach.

In study [70] authors, as preprocessing of Tweets work on identification and removal of bot-generated tweets from social media data streams, employing a refined filter list derived from geo-tagged tweets collected via the Twitter Streaming API. Following preprocessing, supervised-learning algorithms are applied to classify tweets into domain-relevant and domain-irrelevant categories, facilitating precise sentiment analysis for disaster management purposes. For sentiment classification are used Naive Bayes (NB), SVM, Random Decision Forest (RF) and Feedforward Neural Networks (FNNs).

For instance, in the paper [71] on sentiment analysis of forest fires, the TextBlob model, an open-source Python library, was used for the analysis. Besides sentiment analysis, the library offers a consistent API for various NLP tasks like part-of-speech tagging and noun phrase extraction, using a predefined dictionary to classify words as positive or negative for sentiment determination. However, the researchers concluded that the model could not accurately assess sentiment due to the presence of too many irrelevant tweets.

Authors of paper [72] did research of how presence of zip codes in Tweets affects the performance of machine learning models by simulating different percentages of positive samples labeled with zip codes. Classification algorithms that were used in this research are Logistic Regression, K-Nearest Neighbours, NB, RF, SVM, Long Short-Term Memory

(LSTM), Word Embedded Convolutional Neural Network (CNN). The results represents that all machine learning classifiers, with the exception of KNN and Naïve Bayes, achieve state-of-the-art performance in detecting rescue requests from social media.

Paper [73] introduces multi-level framework for situational awareness using Twitter data. They evaluated the effectiveness of cross-events Tweets for perception-level and humanitarian-level situational awareness, addressing the limitations of relying simply on event-specific dataset. Additionally, the accuracy of a dependency parser was tested, to ensure reliable extraction of actionable noun-verb pairs for action-level situational awareness detection. SVM, CNN, LR are popular machine learning methods used to classify the relevancy of SM posts but they can't give direct disaster information.

So authors of paper [74] developed a Large Language model-based approach that takes advantage of its strong language comprehension and few shot learning capabilities. This model is based on Hierarchical Event Classifier and few-shot learning approach.

Paper [75] is using lexicon-based approach for sentiment research. This approach is useful for texts that lack fine detail or specificity.

There is also algorithm developed by MonkeyLearn that can help in identifying polarity in the textual data. Combination of that MonkeyLearn algorithm and supervised classification of sentiment was obtained in paper [76]. MonkeyLearn [77] is a Machine Learning platform for Text Analysis, helping users extract actionable data from raw text. It detects topics or sentiments in texts like tweets, reviews, and articles. MonkeyLearn models are categorized into two types: classification, which assigns labels to text, and extraction, which identifies specific data within text.

	SVM	CNN	LR	NB	Multinominal NB	XGB
[23]					x	x
[36]					x	x
[68]	x					
[73]	x	x	x			
[75]	x			x		
[78]				x		

Table 6.2: Techniques used for classification

Table 6.2 presents various classification techniques (SVM, CNN, LR, NB, Multinomial NB, and XGB) used in different studies.

6.3. Social Media Data Integration for Enhanced Data Analysis

Deep learning (DL) [79], a branch of machine learning, uses neural networks to process data through multiple layers of artificial neurons. Inspired by biological neuroscience, these networks can be trained in supervised, semi-supervised, or unsupervised ways. Common architectures include fully connected networks, convolutional and recurrent neural networks, deep belief networks, GANs, transformers, and neural radiance fields. In machine learning, a neural network, also known as an artificial neural network (ANN) or neural net, is a computational model designed to mimic the structure and workings of biological neural networks found in animal brains [80]. Recurrent neural networks (RNNs) are a type of artificial neural network designed for handling sequential data. RNNs work through multiple time steps, which makes them particularly suited for tasks involving text, speech, and time series analysis [81]. RNNs are used in Natural Language Processing to capture sequential information, unlike traditional neural networks, where inputs are treated independently. For tasks like predicting the next word in a sentence, knowing the previous word is crucial. RNNs process each element of a sequence by considering both the current input and prior outputs, making them capable of handling context. However, they are limited in how far back they can look, which is why LSTM units are added. LSTMs act like memory units, allowing RNNs to retain context from earlier in the sequence for more accurate predictions [82]. The LSTM architecture is made up of recurrently connected units called memory blocks. These blocks are designed to preserve their internal state over time and control the flow of information using non-linear gating mechanisms [83]. BiLSTM is an highly developed form of RNNs. It works well with textual data from social media [84]. In NLP, deep learning has enabled the development of advanced models that excel at understanding and generating human language. One prominent model is BERT (Bidirectional Encoder Representations from Transformers), which uses transformers to process language bidirectionally, capturing context from both preceding and succeeding words [85]. These transformers form the backbone of larger language models, such as GPT, that have demonstrated remarkable capabilities in language generation, translation, and comprehension across various domains. The rapid expansion of Big Data, however, drives the need for more scalable and commercially viable deep learning (DL) algorithms for data analytics and knowledge discovery. As a result, the modern ecosystems of Big Data and DL are closely interconnected, with shared applications and a mutual influence on each other's progress [22].

The increasing frequency and severity of natural disasters highlight the need for advanced technologies to provide timely and accurate assessments of their impacts. Both textual and image data have proven to be valuable resources for disaster monitoring and management, yet each modality has its limitations. Recent advancements in artificial intelligence (AI) and machine learning (ML) have opened up new possibilities for integrating textual and image data to enhance disaster response efforts. The combination of these data types enables a more comprehensive understanding of disaster events by leveraging the strengths of both modalities. . By discussing models that range from computer vision to natural language processing (NLP), it will be demonstrated the potential of combining multimodal data sources for disaster mitigation and recovery.

Textual data as itself has their advantages and disadvantages. It is noisy with a lot of useless information that needs to be preprocessed as it was described in previous section. Satellite images often do not have a high enough resolution for detailed analysis at ground level, which can limit the possibility of precise damage assessment. Processing and analysis of satellite imagery can take time, which can lead to delays in providing key information. But, textual and image data together can give very useful information about natural disaster. Text data from social networks can provide immediate information from witnesses and affected people, while satellite images can provide a broader overview of the situation on the ground. Information from social networks can be verified and confirmed through satellite images, reducing the risk of disinformation spreading. While satellite imagery allows visualization of the scale of a disaster, textual data can provide details about specific places, conditions and population needs. To monitor phase transitions during flood caused by Hurricane Harvey in the Houston area, authors of paper [86] train a computer vision model to classify social media images into four disaster phases: preparedness, impact, response, and recovery. To identify emergency incidents, they use a deep learning NLP model to detect locations from tweet texts, assigning geographic coordinates using a local gazetteer based on GeoNames and TIGER US Census data because GeoNames doesn't provide road information. By combining image and text analysis, they filter tweets with "Impact" images and precise locations for valuable information. A manual review enhances the AI results, aiding comprehensive situation awareness and supporting rescue operations. Paper [6] analyze the potential of estimating disaster impacts through the analysis of Twitter posts before, during and after hurricane Uri in Texas. With such data they automatically calculated keywords connected to Uri. Their method can yield actionable insights to aid first responders, volunteers, governments, and the general public. Prototype utilizes a deep neural network (DNN) to analyze geolocated social media posts (including text, images, and videos), providing monetary damage assessments at the zip code level for the effects of Uri, achieving up to 70% accuracy. For event detection, authors of paper [87] propose algorithm that integrates text, image and statistical information. After semantic and statistical analysis, tweets group together in clusters and then image analysis is performed to separate events from non-events. In paper [88] authors use textual data and visual data for flood detection. They propose the Deep Social media Data Classification (DeepSDC). Textual data is processed using RoBERT and XLNet and images are processed with pretrained networks VGG16 and ResNet50. DeepSDC use augmentation, up-sampling and down-sampling techniques.

Paper [89] introduces multi-dimensional impact assessment of floods by combining multiple sources of data. As main sources of data authors use: Schneider Electric (SE), High-resolution multispectral satellite imagery data (PlanetScope product), Orange - Call Detail Records and Crimson Hexagon (CH). His research demonstrated that integrating social media with satellite imagery, environmental data, and real-time location information from Call Detail Records could improve existing disaster management practices and offer a comprehensive evaluation of the impact.

Author of paper [90] use social media Facebook and Twitter as channel for dissemination of graphical bushfire content to rural Australians. Idea of the paper is to create map of bushfire using GIS and satellite image from MODIS (Moderate-resolution Imaging Spectroradiometer) system. After dissemination of created map, goal is to share it through social media and to attract attention of rural communities to participate in bushfire mitigation efforts. The combination of social media data, meteorological data, and satellite imagery can help predict the behavior of a disaster and its consequences. For example, changes in temperature, atmospheric conditions, or cloud movements can be combined with information from social media to identify potential hazards.

Furthermore, the study [91] discusses the combination of data from newspapers, satellite imagery, social media, and meteorological data. There were free available datasets for fires in Indonesia, so authors used scrapping GetOldTweets-python script to get data. Tweets do not include GPS langitude and longitude position, due to security reasons so that part was solved with using TAGGS algorithm for geoparsing tweets based on the tweet content. Twitter data about forest fires in Riau for period of four years was compared with MODIS satellite images. In results of plots there are consistency between them. When number of hotspots are large, the number of posted tweets is also high, especially during dry season.

In study [92], typhoon event information extraction was approached as a sequence labeling problem. The BiLSTM-CRF model, typically used for Named Entity Recognition, was enhanced with a BERT pre-trained language model for word vector embedding. The process involved inputting character sequences into BERT, which generated word vectors with contextual semantic information. These vectors were then processed by the BiLSTM layer, where the forward LSTM learned future features and the backward LSTM learned past features. The BiLSTM outputs were decoded by Conditional Random Field (CRF) to produce the optimal labeling sequence.

The study [84] introduces a novel approach for event detection, integrating active learning with BiLSTM. Active learning, a semi-supervised technique, enhances the model training process by dynamically selecting the most informative data for annotation. The integration of Bi-LSTM with active learning for fast disaster detection works in a cyclical process. Initially, the Bi-LSTM model is trained on a small labeled dataset, and through active learning, it identifies ambiguous or informative posts from an unlabeled dataset for manual annotation. As the model is retrained with newly labeled data, it continuously improves its ability to detect disaster-related information in real-time social media posts.

Large Language Models

Transformer is type of deep learning architecture composed of an encoder and a decoder, represented in 2017 in paper [93]. This architecture took NLP to a higher level. It is completely new approach for NLP tasks that entirely utilizes the self-attention mechanism. That means that mechanism enables the model to assess the significance of various words and in a sentence when encoding it into a fixed-size vector representation. This leads to faster training time such as parallelization on GPUs. Transformer encoder and decoder are made of multiple layers of self-attention and feedforward neural networks. Encoder-decoder architecture is usually used for language translations. As input model takes sentence and encodes it into fixed-size vector representation. After that decoders job is to generate the output sentence in a way that uses both self-attention and cross-attention. Text-to-text Transfer Transformer T5 is most popular encoder-decoder model. Encoder model can be used for text classification. In that case model must recognize predefined text category. Encoder model is focused only on input values and it doesn't generate any output sequence. Decoder architecture gives output sequences based on input fixed-size vector representation. One of the most popular Transformer decoder model is GPT-3. Transformer architecture was developed to address the challenges of converting one sequence of data to another.

One of the most popular Transformer encoder model is BERT. It is a pre-trained language model developed by Google that uses a transformer architecture to understand the context of words in both directions (left-to-right and right-to-left). BERT captures relationships between words and sub-words (word embeddings) to encode text. BERT is employed in this study [94] to address the challenge of classifying tweets that may use the word "flood" in both literal and figurative senses across multiple languages. As tweets do not always refer to ongoing flood events but may also reference historical or predicted events, BERT is used to accurately distinguish between tweets relevant to current events and those that are not. By leveraging BERT's deep learning capabilities, which encode word and sub-word relationships through embeddings, the model can capture subtle semantic differences, making it suitable for multilingual text classification. The pre-trained BERT model is fine-tuned on a manually labeled dataset of tweets, ensuring it can be adapted to the specific task of filtering tweets for event relevance across different languages. This approach is particularly beneficial for handling spatial and temporal variations in the data, as the classification process ensures that tweets referring to ongoing events are properly identified and separated from unrelated or metaphorical uses.

RoBERTa (Robustly optimized BERT approach) is an NLP model developed by Facebook AI, designed to improve upon BERT (Bidirectional Encoder Representations from Transformers). RoBERTa was trained on a significantly larger dataset compared to BERT (160 GB of textual data from books, news, stories and Wikipedia) [88]. It is used in paper [88] for text analysis. Authors of paper [95] reveals a gap in literature in the absence of pre-trained language models designed specifically for crisis-related social media content. Although domain-specific models have shown success across various fields, they do not account for the distinct language used during crises. In response, the authors introduce CrisisTransformers, a collection of pre-trained models and sentence encoders built on extensive crisis-related social media data, aimed at improving the analysis of these texts. Authors of paper [96] incorporated typhoon disaster knowledge into the T5 model, utilizing a Q&A framework to deliver high-quality information effectively. The T5 model, or "text-to-text transfer transformer," utilizes a transformer-based encoder-decoder architecture, replacing traditional RNNs or CNNs with self-attention layers for handling input sequences. It converts all natural language processing tasks into a text-to-text format, allowing a single model to address various tasks. By enabling parallel computation and detecting long-range token dependencies, T5 offers efficient performance, especially in text generation. Additionally, its simplified positional embeddings improve processing efficiency across layers.

Deep learning and machine learning techniques, particularly in NLP and computer vision, have shown immense potential in enhancing disaster response by integrating textual and image data. By combining social media posts with satellite imagery, these models can provide a more accurate and comprehensive assessment of disaster impacts, improving situational awareness and aiding emergency response efforts. Advanced models like BERT, RoBERTa, and Crisis-Transformers further enhance the ability to process crisis-related data efficiently, supporting timely and effective disaster management.

6.4. Time series

A time series is a sequence of data points collected or recorded at specific time intervals, often with uniform spacing. If data is observed through certain time then it can be called time series data. Behavior of data through time can be observed with time series analysis. Time series analysis usually has one of these two goals: to predict next value or to classify the series. For prediction of next value next models are used: AutoRegressive model (AR), Moving Average (MA), Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA), Long-Short Term Memory (LSTM), Prophet. For classification: extracting features such as mean, median, deviation, skewness, etc. By analysing behavior of data event detection can be established [97]. Time series can reveal in patterns in behavior of natural disaster data, indicate on anomalies, extreme values or highlight changes in the probability distribution.

6.4.1. Time series prediction models

Predicting the next value in a time series has been the focus of many disciplines. Traditional techniques for time series include the Auto Regressive model (AR), Moving Average (MA), and enhancements such as Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) [98]. However, in the context of time series analysis of social media posts for disaster management, there is no literature leveraging these methods.

Long-short term memory (LSTM)

Long Short-Term Memory (LSTM) networks are widely used in time series analysis, particularly for tasks involving sequential data. LSTM networks, a special type of RNN, can model long-term relationships between input and output data using feedback mechanisms and gates that retain past information. This allows the model to learn from both past and present data. Research shows that deep learning models, like LSTM, outperform traditional ARIMA models in time series forecasting, especially for long-term predictions [99]. The main advantage of LSTMs is their ability to retain information from long sequences without the problem of vanishing or exploding gradients, which is common in traditional RNNs. This makes LSTMs especially useful in time series problems where long-term dependencies are important.

Prophet

Prophet is an open-source forecasting tool developed by Facebook (now Meta) designed for time series forecasting, particularly suited for handling seasonality, trends, and outliers. It is widely used for business and data science applications due to its flexibility and ease of use. Prophet is particularly effective for time series with strong seasonal patterns and large amounts of missing data or irregular observations [100]. Prophet is designed for non-experts in forecasting and provides simple APIs for Python and R, making it highly accessible for business users and data scientists.

6.4.2. Time series classification

Besides predicting the next value, time series analysis plays a key role in classifying time series data and detecting the specific events they represent. Feature extraction is crucial in this context. Typically, statistical measures derived from the time series are used as features for classification. Table 6.3 provides a summary of key statistical features commonly used for this purpose, including measures like arithmetic mean, standard deviation, skewness, and kurtosis.

Measure	Definition	Equation
Arithmetic mean	Often called the average, it is determined by adding together all the data points in a set and then dividing the total by the number of data points.	$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$(6.1) where x_i represents each data point from 1 to n, and n is the total number of data points [101].
Standard deviation	Widely used metric for assessing the variability within a set of data points [102].	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$(6.2) It is calculated as the square root of the sum of the squared differences between each data point (x_i) and the arithmetic mean ($\bar{X}$), divided by $n-1$ [101].
Skewness	Skewness describes the extent of asymmetry in a distribution or dataset. In a symmetric distribution, both sides appear identical when reflected across the central point. The skewness measure indicates how much a set of values diverges from symmetry around the mean. A normal distribution has a skewness of zero, and symmetric datasets tend to have skewness values near zero. Skewness can be positive, meaning the distribution is right-skewed with a longer and thinner right tail, or negative, indicating left-skewness where the left tail is longer and thinner. Data may become skewed due to constraints, such as in reliability studies where failure times can't be negative, often leading to right-skewed distributions.	The skewness coefficient for univariate data Y_t is defined as: $S = \frac{1}{(n\sigma)^3}\sum_{t=1}^{n}(Y_t\overline{Y}_t)^3$(6.3) where $\overline{Y}_t$ stands for mean, σ for standard deviation and n for number of data points [103].
Kurtosis	Kurtosis quantifies how much a dataset differs from a normal distribution based on its peak and tail behavior. A dataset with high kurtosis has a pronounced peak near the mean and rapidly tapering tails. In contrast, low kurtosis indicates a flatter, more spread-out top around the mean without a sharp peak. High kurtosis also suggests heavy tails, meaning extreme values are more frequent compared to a normal distribution. Conversely, low kurtosis indicates lighter tails, making extreme values less likely to appear.	Kurtosis coefficient for a univariate time series Y_t is defined as: $K = \frac{1}{(n\sigma)^4}\sum_{t=1}^{n}(Y_t\overline{Y}_t)^4$(6.4) where $\overline{Y}_t$ stands for mean, σ for standard deviation and n for number of data points [103].

Table 6.3: Time series classification features with definitions and equations.

In general, in the entire reviewed literature, there is a lack of works related to the analysis of time series from the texts of social networks. That's exactly why we dealt with it in the previous work. Thus, the paper [104] compared data from official sources and social networks using time series analysis to obtain data correlation, which resulted in certain overlaps and thus concluded that social network data is a very important source of data for such analyses. However, further research is needed in this direction to refine the methodologies for extracting meaningful patterns from social media data. Additionally, integrating real-time analysis with predictive modeling could enhance the effectiveness of disaster management systems by providing timely insights.

7. CONCLUSION

This document summarizes the relevant scientific literature on state-of-the-art approaches for utilizing user-generated social media texts as valuable information in disaster management. However, such advancements cannot be achieved without reliable observations. During and after natural disasters, many researchers rely on in-situ sampling and measurements, which are both costly and time-consuming. In contrast, besides government and civil protection bodies, there are often numerous active and passive observers present at disaster sites. Leveraging the efforts of these observers could enhance our understanding of natural disasters, though several obstacles must first be overcome. While the existing research on employing social media for disaster mitigation is thorough, certain limitations persist. For instance, there is limited research focused on assessing the magnitude of natural disasters. Challenge for further work can be that most people in our region uses Facebook as main social network. Lot of research put focus on different impacts after natural disaster. For example impact on human settlements, enivornment, economy, turism, agricultural production, housing, schools, etc. SISM (Social Impact in Social Media) methodology is used in social media analytics and as a methodology to evaluate the social impact of research [55]. To answer research questions that were set up at the beginning of this paper 57 papers were included.

What crisis situation-natural and other disasters are considered by social media? What types of disasters are mostly covered in literature and in what extent?

The literature on disaster research encompasses a broad spectrum of crisis situations, including both natural but also and other types of disasters. In this document, focus is on natural disasters such as hurricanes, typhoons, floods, earthquakes, and wildfires. Each of these events presents unique challenges and impacts on affected communities, making them critical areas of study. Additionally, the literature also addresses significant global events such as the COVID-19 pandemic and terrorist attacks, which, while not classified strictly as natural disasters, have profound implications for public health, safety, and social dynamics.

For the purpose of this research, the focus is primarily on natural disasters. Among the types analyzed, floods are the most frequently mentioned and have been identified as the most common natural disaster, as elaborated in Section 4. Floods not only occur with high frequency but also result in considerable social, economic, and environmental impacts, which makes them a focal point for many studies. It is important to note that the frequency of natural disaster types in the relevant literature correlates with their occurrence. Thus, we can confidently conclude that the impact of natural disasters on social media is studied in line with their relevance—more frequent disasters are examined more often.

What approaches are taken to code user generated data into quantities representation (impact assessment)?

In view of approaches we can conclude that each research in text analysis has three main steps: (1) data preparation , (2) vectorization and (3) final processing. All of these steps are described in Section 6. Different type of preprocessing text like keywords extraction/filtering, normalization, tokenization leads to vectorization techniques that helps turn text into structured representations that can be processed with algorithms like SVM, Naive Bayes, etc. In reviewed literature there is lack of papers that give importance on impact factor in meaning that factor gives numerical description how tough was some natural disaster. By impact factor, we mean a specific weighting coefficient that would include the calculation of how many people, organizations, emergency services participated during the intervention, how much area was covered, and so on.

What techniques are used to transform the user generated data into predictions?

Mostly classification time series analysis are described in Section 6 also. They can be divided in two groups: to predict next value or to classify the series. In the context of predicting the next value, a range of techniques has been developed, including the AutoRegressive (AR) model, Moving Average (MA), and more sophisticated methods like the Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM) and Propeth. For classification, the focus is on feature extraction, utilizing statistics such as mean, standard deviation, and skewness to evaluate changes in data behavior over time. In literature that is included in this paper there is lack of this kind of analysis of natural disaster. All papers are focused on one certain event that happened in one short period of time and there is gap in research of natural disasters in some longer period of that where patterns, anomalies can be detected using time series analysis.

In summary, while significant steps forward have been made in utilizing social media for disaster research, there are critical gaps that need further research. Future studies should aim to develop standardized impact assessment metrics, especially focusing on quantifying the severity of natural disasters. Additionally, there is a pressing need to conduct longitudinal studies employing time-series analysis to detect patterns and anomalies over time. Addressing these gaps will not only enhance our understanding of disaster dynamics but also improve preparedness and response strategies for future events.

LITERATURE

[1] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown et al., Tackling climate change with machine learning, *ACM Computing Surveys (CSUR)*, 55, 2, 1–96, 2022.

[2] D. Adegoke, A systematic review of big data and digital technologies security leadership outcomes effectiveness during natural disasters, *Sustainable Futures*, 100113, 2023.

[3] A. Rudyawan, A. Pamumpuni, M. Puspaningrum, I. Gunawan, B. Sapiie, A. Handayani i M. Abdurrachman, Cianjur earthquake-roles of social media and the distribution of sciences, *IOP Conference Series: Earth and Environmental Science*, 1245, 012016, IOP Publishing, 2023.

[4] Twitter, "twitter", `https://x.com/`, Accessed: 15/05/2024.

[5] Fb, "facebook", `https://www.facebook.com/`, Accessed: 15/05/2024.

[6] Y. Pi, X. Ye, N. Duffield i on behalf of the Microsoft AI for Humanitarian Action Group, Rapid damage estimation of texas winter storm uri from social media using deep neural networks, *Urban Science*, 6, 3, 2022.

[7] M. J. Aramburu, R. Berlanga i I. Lanza-Cruz, A data quality multidimensional model for social media analysis, *Business & Information Systems Engineering*, 1–23, 2023.

[8] L. Šerić, A. Ivanda i S. Knežić Buhovac, Towards an integrated framework for collaborative modeling using social media posts and earth observations, *Big Data from space 2023*, 297–300, 2023.

[9] L. Shamseer, D. Moher, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle i L. A. Stewart, Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: elaboration and explanation, *Bmj*, 349, 2015.

[10] S. A. Khan, K. Shahzad, O. Shabbir i A. Iqbal, Developing a framework for fake news diffusion control (fndc) on digital media (dm): A systematic review 2010–2022, *Sustainability*, 14, 22, 2022.

[11] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan et al., The prisma 2020 statement: an updated guideline for reporting systematic reviews, *International journal of surgery*, 88, 105906, 2021.

[12] Q. Hou, M. Han i Z. Cai, Survey on data analysis in social media: A practical application aspect, *Big Data Mining and Analytics*, 3, 4, 259–279, 2020.

[13] U. A. Bukar, F. Sidi, M. A. Jabar, R. N. H. Nor, S. Abdullah, I. Ishak, M. Alabadla i A. Alkhalifah, How advanced technological approaches are reshaping sustainable social media crisis management and communication: a systematic review, *Sustainability*, 14, 10, 5854, 2022.

[14] D. Contreras, S. Wilkinson i P. James, Earthquake reconnaissance data sources, a literature review, *Earth*, 2, 4, 1006–1037, 2021.

[15] Y. Wang, J. Li, X. Zhao, G. Feng i X. Luo, Using mobile phone data for emergency management: a systematic literature review, *Information Systems Frontiers*, 22, 1539–1559, 2020.

[16] W. Su, L. Chen i X. Gao, Emergency decision making: A literature review and future directions, *Sustainability*, 14, 17, 10925, 2022.

[17] M. Aboualola, K. Abualsaud, T. Khattab, N. Zorba i H. S. Hassanein, Edge technologies for disaster management: A survey of social media and artificial intelligence integration, *IEEE Access*, 2023.

[18] S. K. Abid, N. Sulaiman, S. W. Chan, U. Nazir, M. Abid, H. Han, A. Ariza-Montes i A. Vega-Muñoz, Toward an integrated disaster management approach: how artificial intelligence can boost disaster management, *Sustainability*, 13, 22, 12560, 2021.

[19] A. Tounsi i M. Temimi, A systematic review of natural language processing applications for hydrometeorological hazards assessment, *Natural hazards*, 116, 3, 2819–2870, 2023.

[20] C. Zhang, C. Fan, W. Yao, X. Hu i A. Mostafavi, Social media for intelligent public information and warning in disasters: An interdisciplinary review, *International Journal of Information Management*, 49, 190–207, 2019.

[21] M. Krichen, M. S. Abdalzaher, M. Elwekeil i M. M. Fouda, Managing natural disasters: An analysis of technological advancements, opportunities, and challenges, *Internet of Things and Cyber-Physical Systems*, 2023.

[22] A. Y. Sun i B. R. Scanlon, How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions, *Environmental Research Letters*, 14, 7, 073001, 2019.

[23] M. Raza, M. Awais, K. Ali, N. Aslam, V. V. Paranthaman, M. Imran i F. Ali, Establishing effective communications in disaster affected areas and artificial intelligence based detection using social media platform, *Future Generation Computer Systems*, 112, 1057–1069, 2020.

[24] Em-dat, "natural disaster in numbers", `https://www.emdat.be/`, Accessed: 15/07/2024.

[25] A. Vavassori, D. Carrion, B. Zaragozi i F. Migliaccio, Vgi and satellite imagery integration for crisis mapping of flood events, *ISPRS International Journal of Geo-Information*, 11, 12, 2022.

[26] J. Gao, O. Murao, X. Pei i Y. Dong, Identifying evacuation needs and resources based on volunteered geographic information: a case of the rainstorm in july 2021, zhengzhou, china, *International journal of environmental research and public health*, 19, 23, 16051, 2022.

[27] Dyfi, "did you feel it?", 1999, https://earthquake.usgs.gov/data/dyfi/ [Accessed: (10/7/2024)].

[28] Emsc, "lastquake", https://m.emsc.eu/ [Accessed: (10/7/2024)].

[29] Ems, "copernicus emergency management service", https://emergency.copernicus.eu/, Accessed: 10/07/2024.

[30] Fcs, "floodcitisense", http://www.floodcitisense.eu/ [Accessed: (10/7/2024)].

[31] Shimo, "shimo", https://shimo.im/, Accessed: 01/09/2024.

[32] Weibo, "weibo", https://weibo.com/, Accessed: 15/05/2024.

[33] Wechat, "wechat", https://www.wechat.com/, Accessed: 01/09/2024.

[34] T. Tang i G. Hu, Detecting and tracking significant events for individuals on twitter by monitoring the evolution of twitter followership networks, *Information*, 11, 9, 450, 2020.

[35] Domo "data never sleeps", 2021, https://www.domo.com/data-never-sleeps#timeline [Accessed: (21/5/2024)].

[36] Z. Yang i Q. Wang, Mining method of traffic impact areas of rainstorm event based on social media in zhengzhou city, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 79–86, 2022.

[37] T. Yigitcanlar, M. Regona, N. Kankanamge, R. Mehmood, J. D'Costa, S. Lindsay, S. Nelson i A. Brhane, Detecting natural hazard-related disaster impacts with social media analytics: the case of australian states and territories, *Sustainability*, 14, 2, 810, 2022.

[38] Twitter: Academic research product track", 2021, https://devcommunity.x.com/t/introducing-the-new-academic-research-product-track/148632 [Accessed: (1/6/2024)].

[39] Twitterapichanges, "twitterapi", https://medium.com/@asaan/twitter-api-changes-navigating-the-end-of-free-access-your-2024-guide-b9f9cf47ea79, [Accessed: (23/8/2024)].

[40] Ch, "crimson hexagon", https://www.brandwatch.com/, Accessed: 15/05/2024.

[41] C. Fan i A. Mostafavi, Metanetwork framework for performance analysis of disaster management system-of-systems, *IEEE Systems Journal*, 14, 1, 1265–1276, 2020.

[42] C. E. S. Kuhn, F. A. G. V. Reis, A. F. Lazaretti, C. Zarfl, V. C. Cabral i P. Grathwohl, The record and trends of natural disasters caused by gullies in brazil, *Environmental Earth Sciences*, 82, 22, 524, 2023.

[43] R. Verma, S. Karimi, D. Lee, O. Gnawali i A. Shakery, Newswire versus social media for disaster response and recovery, *2019 Resilience Week (RWS)*, 1, 132–141, 2019.

[44] S. Pant, E. N. Yadav, Milan, M. Sharma, Y. Bedi i A. Raturi, Web scraping using beautiful soup, *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, 1, 1–6, 2024.

[45] C. Lotfi, S. Srinivasan, M. Ertz i I. Latrous, Web scraping techniques and applications: A literature.

[46] V. Morina i S. Sejdiu, Evaluating and comparing web scraping tools and techniques for data collection, *UBT Knowl. Cent*, 2022.

[47] A. G. Gulnerman, H. Karaman, D. Pekaslan i S. Bilgi, Citizens' spatial footprint on twitter—anomaly, trend and bias investigation in istanbul, *ISPRS International Journal of Geo-Information*, 9, 4, 222, 2020.

[48] M. Tavra, I. Racetin i J. Peroš, The role of crowdsourcing and social media in crisis mapping: a case study of a wildfire reaching croatian city of split, *Geoenvironmental Disasters*, 8, 1–16, 2021.

[49] M. Mirbabaie, D. Bunker, S. Stieglitz, J. Marx i C. Ehnis, Social media in times of crisis: Learning from hurricane harvey for the coronavirus disease 2019 pandemic response, *Journal of Information Technology*, 35, 3, 195–213, 2020.

[50] S. Beliga, A. Meštrović i S. Martinčić-Ipšić, An overview of graph-based keyword extraction methods and approaches, *Journal of information and organizational sciences*, 39, 1, 1–20, 2015.

[51] S. Kiran, C. R. S. Sai, M. Pooja et al., A comparative study on parsing in natural language processing, *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 785–788, IEEE, 2019.

[52] A. Backurs i P. Indyk, Which regular expression patterns are hard to match?, *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 457–466, IEEE, 2016.

[53] D. H. Maulud, S. Y. Ameen, N. Omar, S. F. Kak, Z. N. Rashid, H. M. Yasin, I. M. Ibrahim, A. A. Salih, N. O. Salim i D. M. Ahmed, Review on natural language processing based on different techniques, *Asian Journal of Research in Computer Science*, 10, 1, 1–17, 2021.

[54] D. Khyani, B. Siddhartha, N. Niveditha i B. Divya, An interpretation of lemmatization and stemming in natural language processing, *Journal of University of Shanghai for Science and Technology*, 22, 10, 350–357, 2021.

[55] E. Oliver, M. C. Llasat, M. Llasat-Botija i J. Díez-Palomar, Twitter's messages about hydrometeorological events. a study on the social impact of climate change, *Sustainability*, 13, 6, 3579, 2021.

[56] T. Yang, J. Xie, G. Li, N. Mou, C. Chen, J. Zhao, Z. Liu i Z. Lin, Traffic impact area detection and spatiotemporal influence assessment for disaster reduction based on social media: A case study of the 2018 beijing rainstorm, *ISPRS International Journal of Geo-Information*, 9, 2, 2020.

[57] U. Krzeszewska, A. Poniszewska-Marańda i J. Ochelska-Mierzejewska, Systematic comparison of vectorization methods in classification context, *Applied Sciences*, 12, 10, 5119, 2022.

[58] S. Dai, K. Li, Z. Luo, P. Zhao, B. Hong, A. Zhu i J. Liu, Ai-based nlp section discusses the application and effect of bag-of-words models and tf-idf in nlp tasks, *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 5, 1, 13–21, 2024.

[59] D. Jurafsky i J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 2024, online manuscript released August 20, 2024.

[60] P. Bruneau, E. Brangbour, S. Marchand-Maillet, R. Hostache, M. Chini, R.-M. Pelich, P. Matgen i T. Tamisier, Measuring the impact of natural hazards with citizen science: The case of flooded area estimation using twitter, *Remote Sensing*, 13, 6, 1153, 2021.

[61] D. M. Blei, A. Y. Ng i M. I. Jordan, Latent dirichlet allocation, *Journal of machine Learning research*, 3, Jan, 993–1022, 2003.

[62] V. V. Mihunov, N. H. Jafari, K. Wang, N. S. Lam i D. Govender, Disaster impacts surveillance from social media with topic modeling and feature extraction: case of hurricane harvey, *International Journal of Disaster Risk Science*, 13, 5, 729–742, 2022.

[63] H. Safarpour i L. A. Spearing, Temporal public perceptions and experiences during water service disruptions: the case of jackson, mississippi, *Environmental Research: Infrastructure and Sustainability*, 4, 1, 015003, 2024.

[64] J. Durham, S. Chowdhury i A. Alzarrad, Unveiling key themes and establishing a hierarchical taxonomy of disaster-related tweets: A text mining approach for enhanced emergency management planning, *Information*, 14, 7, 2023.

[65] H. Zhu i K. Liu, Temporal, spatial, and socioeconomic dynamics in social media thematic emphases during typhoon mangkhut, *Sustainability*, 13, 13, 2021.

[66] A. Gasparetto, M. Marcuzzo, A. Zangari i A. Albarelli, A survey on text classification algorithms: From text to predictions, *Information*, 13, 2, 83, 2022.

[67] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes i D. Brown, Text classification algorithms: A survey, *Information*, 10, 4, 150, 2019.

[68] S. Ahadzadeh i M. R. Malek, Earthquake damage assessment based on user generated data in social networks, *Sustainability*, 13, 9, 2021.

[69] J. Kersten i F. Klan, What happens where during disasters? a workflow for the multifaceted characterization of crisis events based on twitter data, *Journal of Contingencies and Crisis Management*, 28, 3, 262–280, 2020.

[70] F. Yao i Y. Wang, Domain-specific sentiment analysis for tweets during hurricanes (dssa-h): A domain-adversarial neural-network-based approach, *Computers, Environment and Urban Systems*, 83, 101522, 2020.

[71] K. Milleville, S. Van Ackere, J. Verdoodt, S. Verstockt, P. De Maeyer i N. Van de Weghe, Exploring the potential of social media to study environmental topics and natural disasters, *Journal of Location Based Services*, 17, 4, 355–369, 2023.

[72] Z. Wang, N. S. Lam, M. Sun, X. Huang, J. Shang, L. Zou, Y. Wu i V. V. Mihunov, A machine learning approach for detecting rescue requests from social media, *ISPRS International Journal of Geo-Information*, 11, 11, 570, 2022.

[73] W. Zhai, A multi-level analytic framework for disaster situational awareness using twitter data, *Computational Urban Science*, 2, 1, 23, 2022.

[74] J. Hou i S. Xu, Near-real-time seismic human fatality information retrieval from social media with few-shot large-language models, *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 1141–1147, 2022.

[75] S. Lu, J. Huang i J. Wu, Multi-dimensional urban flooding impact assessment leveraging social media data: A case study of the 2020 guangzhou rainstorm, *Water*, 15, 24, 4296, 2023.

[76] D. Contreras, S. Wilkinson, E. Alterman i J. Hervás, Accuracy of a pre-trained sentiment analysis (sa) classification model on tweets related to emergency response and early recovery assessment: the case of 2019 albanian earthquake, *Natural Hazards*, 113, 1, 403–421, 2022.

[77] Monkeylearn, "monkeylearn", `https://help.monkeylearn.com/en/`, Accessed: 04/09/2024.

[78] F. P. Salsabilla i D. R. Hizbaron, Understanding community collective behaviour through social media responses: Case of sunda strait tsunami, 2018, indonesia, *E3S Web of Conferences*, 325, 01021, EDP Sciences, 2021.

[79] Y. LeCun, Y. Bengio i G. Hinton, Deep learning, *nature*, 521, 7553, 436–444, 2015.

[80] S. C. Peter, J. K. Dhanjal, V. Malik, N. Radhakrishnan, M. Jayakanthan i D. Sundar, Quantitative structure-activity relationship (qsar): Modeling approaches to biological applications, S. Ranganathan, M. Gribskov, K. Nakai i C. Schönbach, editori, *Encyclopedia of Bioinformatics and Computational Biology*, 661–676, Academic Press, Oxford, 2019.

[81] A. Tealab, Time series forecasting using artificial neural networks methodologies: A systematic review, *Future Computing and Informatics Journal*, 3, 2, 334–340, 2018.

[82] K. M. Tarwani i S. Edem, Survey on recurrent neural network in natural language processing, *Int. J. Eng. Trends Technol*, 48, 6, 301–304, 2017.

[83] G. Van Houdt, C. Mosquera i G. Nápoles, A review on the long short-term memory model, *Artificial Intelligence Review*, 53, 8, 5929–5955, 2020.

[84] P. Adili i Y. Chen, Fast disaster event detection from social media: An active learning method, *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*, 19, 2, 2024.

[85] J. D. M.-W. C. Kenton i L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of naacL-HLT*, 1, 2, 2019.

[86] R.-Q. Wang, Y. Hu, Z. Zhou i K. Yang, Tracking flooding phase transitions and establishing a passive hotline with ai-enabled social media data, *IEEE Access*, 8, 103395–103404, 2020.

[87] Y. Han, S. Karunasekera i C. Leckie, Image analysis enhanced event detection from geo-tagged tweet streams, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 398–410, Springer, 2020.

[88] M. Hanif, M. Waqas, A. Muneer, A. Alwadain, M. A. Tahir i M. Rafi, Deepsdc: Deep ensemble learner for the classification of social-media flooding events, *Sustainability*, 15, 7, 6049, 2023.

[89] D. Pastor-Escuredo, Y. Torres, M. Martínez-Torres i P. J. Zufiria, Rapid multi-dimensional impact assessment of floods, *Sustainability*, 12, 10, 2020.

[90] K. Perera, R. Tateishi, K. Akihiko i S. Herath, A combined approach of remote sensing, gis, and social media to create and disseminate bushfire warning contents to rural australia, *Earth*, 2, 4, 715–730, 2021.

[91] A. A. Fitriany, P. J. Flatau, K. Khoirunurrofik i N. F. Riama, Assessment on the use of meteorological and social media information for forest fire detection and prediction in riau, indonesia, *Sustainability*, 13, 20, 11188, 2021.

[92] H. Liu, N. Luo i Q. Zhao, Research on the construction of typhoon disaster chain based on chinese web corpus, *Journal of Marine Science and Engineering*, 10, 1, 44, 2022.

[93] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser i I. Polosukhin, Attention is all you need, *Advances in neural information processing systems*, 30, 2017.

[94] J. A. de Bruijn, H. de Moel, B. Jongman, M. C. de Ruiter, J. Wagemaker i J. C. Aerts, A global database of historic and real-time flood events based on social media, *Scientific data*, 6, 1, 311, 2019.

[95] R. Lamsal, M. R. Read i S. Karunasekera, Crisistransformers: Pre-trained language models and sentence encoders for crisis-related social media texts, *Knowledge-Based Systems*, 296, 111916, 2024.

[96] Y. Xia, Y. Huang, Q. Qiu, X. Zhang, L. Miao i Y. Chen, A question and answering service of typhoon disasters based on the t5 large language model, *ISPRS International Journal of Geo-Information*, 13, 5, 165, 2024.

[97] E. Scharwächter i E. Müller, Two-sample testing for event impacts in time series, *Proceedings of the 2020 SIAM International Conference on Data Mining*, 10–18, SIAM, 2020.

[98] M. Zhang, Time series: Autoregressive models ar, ma, arma, arima, *University of Pittsburgh*, 2018.

[99] S. Siami-Namini, N. Tavakoli i A. S. Namin, The performance of lstm and bilstm in forecasting time series, *2019 IEEE International conference on big data (Big Data)*, 3285–3292, IEEE, 2019.

[100] T. Feng, Z. Zheng, J. Xu, M. Liu, M. Li, H. Jia i X. Yu, The comparative analysis of sarima, facebook prophet, and lstm for road traffic injury prediction in northeast china, *Frontiers in public health*, 10, 946563, 2022.

[101] M. N. Martinez i M. J. Bartholomew, What does it "mean"? a review of interpreting and calculating different types of means and standard deviations, *Pharmaceutics*, 9, 2, 2017.

[102] D. Curran-Everett, Explorations in statistics: standard deviations and standard errors, *Advances in Physiology Education*, 32, 3, 203–208, 2008, pMID: 18794241.

[103] X. Wang, K. Smith i R. Hyndman, Characteristic-based clustering for time series data, *Data mining and knowledge Discovery*, 13, 335–364, 2006.

[104] S. K. Buhovac, L. Šerić i A. Ivanda, Community involvement in fire reporting: Time series examination of official reports and social media posts in split and dalmatia county, *2023 8th International Conference on Smart and Sustainable Technologies (SpliTech)*, 1–6, IEEE, 2023.

LIST OF LABELS AND ACRONYMS

AI - *Artificial intelligence*
ANN - *Artificial Neural Network*
API - *Application Programming Interface*
AR - *AutoRegressive model*
ARIMA - *Autoregressive Integrated Moving Average*
ARMA - *Autoregressive Moving Average*
BERT - *Bidirectional Encoder Representations from Transformers*
CH - *Crimson Hexagon*
CNN - *Convolutional Neural Network*
CRFs - *Conditional Random Fields*
DeepSDC - *Deep Social media Data Classification*
DL - *Deep Learning*
DNN - *Deep Neural Network*
DYFI - *"Did you feel it?"*
EDM - *Emergency Decision Making*
EM - *Emergency Management*
EMS - *Emergency Management Service*
EMSC - *European-Mediterranean Seismological Centre*
EWM - *Early Warning Management*
FNNs - *Feedforward Neural Networks*
GTD - *Geo Tweets Downloader*
IoT - *Internet of Things*
KNN - *k-nearest neighbor*
LDA - *Latent Dirichlet Allocation*
LSA - *Latent Semantic Analysis*
LSTM - *Long-Short Term Memory*
MA - *Moving Average*
ML - *Machine Learning*
MODIS - *Moderate-resolution Imaging Spectroradiometer*
NBC - *Naïve Bayes Classifier*
NLP - *Natural Language Processing*
NLTK - *Natural Language Toolkit*
NMF - *Non-Negative Matrix Factorization*
RE - *Regular Expression*
RF - *Random Decision Forest*

RNNs - *Recurent Neural Networks*

RoBERTa - *Robustly optimized BERT approach*

RS - *Remote Sensing*

SE - *Schneider Electric*

SM - *Social Media*

SMAF - *Social Media Analytics Framework*

SNA - *Social Network Analysis*

SVM - *Support Vector Machines*

TSV - *Tab-Separted Values*

USGS - *United States Geological Survey*

VGI - *Volunteered Geographic information*

WoS - *Web of Science*

ABSTRACT

Natural disaster management is a complex task that is continuously studied and refined. In recent years, alongside traditional scientific and expert data collection methods, social media has emerged as a valuable source of information, offering real-time data from public discourse. This paper systematically reviews scientific literature on the use of user-generated text data from social media (SM) in disaster management by analyzing relevant studies across scientific databases. The literature review aims to address three research questions: RQ1) What types of crisis situations—both natural and other disasters—are commonly addressed through SM? RQ2) What approaches are used to transform user-generated data into quantitative representations, such as impact assessments? And RQ3) What methods are applied to convert user-generated data into predictive insights? Our review found that the most frequently studied natural disasters in this context are floods, earthquakes, storms, and wildfires, addressing RQ1 and underscoring the correlation between the focus on these events and their prevalence. For RQ2, we analyzed techniques for handling textual data from various SM platforms, including data extraction, preprocessing, vectorization, classification, and analysis, as discussed in the scientific literature. In addressing RQ3, we found that while classification methods are commonly explored, there is a notable lack of research on prediction techniques and time-series analysis. Despite progress in integrating SM into disaster management research, this review identifies significant gaps, including the need for standardized impact metrics and longitudinal studies to enhance pattern detection and anomaly identification over time.

SAŽETAK

Upravljanje prirodnim katastrofama složen je zadatak koji se kontinuirano proučava i usavršava. Posljednjih godina, uz tradicionalne znanstvene i stručne metode prikupljanja podataka, društveni mediji su se nametnuli kao vrijedan izvor informacija koji nudi podatke iz javnog diskursa u stvarnom vremenu. Ovaj rad sustavno daje pregled znanstvene literature o korištenju korisnički generiranih tekstualnih podataka iz društvenih medija (SM) u upravljanju katastrofama analizirajući relevantne studije u znanstvenim bazama podataka.

Pregled literature ima za cilj odgovoriti na tri istraživačka pitanja: RQ1) Koje se vrste kriznih situacija—i prirodnih i drugih katastrofa—uobičajeno rješavaju kroz SM? RQ2) Koji se pristupi koriste za transformaciju podataka koje generiraju korisnici u kvantitativne prikaze, kao što su procjene utjecaja? I RQ3) Koje se metode primjenjuju za pretvaranje podataka koje su generirali korisnici u prediktivne uvide?

Naš je pregled otkrio da su najčešće proučavane prirodne katastrofe u ovom kontekstu poplave, potresi, oluje i šumski požari, adresirajući RQ1 i naglašavajući korelaciju između fokusa na te događaje i njihove učestalosti. Za pronaći odgovor na RQ2 analizirali smo tehnike za rukovanje tekstualnim podacima s različitih SM platformi, uključujući ekstrakciju podataka, pretprocesiranje, vektorizaciju, klasifikaciju i analizu, kao što se raspravlja u znanstvenoj literaturi. Obraćajući se RQ3, otkrili smo da iako su metode klasifikacije čest predmet istraživanja, postoji značajan nedostatak istraživanja o tehnikama predviđanja i analizi vremenskih nizova.

Unatoč napretku u integraciji SM-a u istraživanje upravljanja katastrofama, ovaj pregled utvrđuje značajne nedostatke, uključujući potrebu za standardiziranim mjerenjima utjecaja i longitudinalnim studijama za poboljšanje otkrivanja uzoraka i identifikacije anomalija tijekom vremena.