

DATA MINING AND SPATIAL ANALYSIS OF TWITTER AS A RESOURCE FOR ASSESSING UK WATER POLLUTION

A. Falaye

11th September 2023

Nottingham Geospatial Institute, University of Nottingham, UK

Dissertation submitted to the University of Nottingham in partial fulfilment of Masters of Research
in Geospatial Systems

Abstract

Water pollution is a significant global menace to human health, ecosystems, and economic progress. Despite advancements towards achieving the United Nations Sustainable Development Goals (SDGs) regarding accessible clean water and sanitation, water pollution remains a substantial hurdle. Active involvement of the public is vital in curbing water pollution, and comprehending their viewpoints and knowledge is pivotal for effective behavior modification. The indispensability of social media in our daily lives cannot be overstated, as it has emerged as the preeminent influential platform for educating society and gathering public perceptions. By utilizing data mining techniques, Social media platforms has the potential to convert public opinions into invaluable Volunteered Geographic Information (VGI), thus promoting citizen science. This study duly recognizes the potency of social media, particularly Twitter, as an efficient tool for mapping water pollution patterns, trends, events, and public sentiments. By conducting spatial analysis and data mining on Twitter data, a wealth of valuable insights is unveiled. These insights encompass the identification of pollution hotspots, pinpointing event locations, the acquisition of knowledge regarding their underlying causes of pollution, regional disparities in these causes, measurement of sentiments on the topic across regions, and the formulation of potential resource management strategies. These contributions are in direct alignment with the overarching goal of achieving a net-zero-plus future, which resonates with the UN SDGs.

Keywords: Volunteered Geographic Information, Social Media, Twitter, Data Mining, Latent Dirichlet Allocation, Water Pollution.

Contents

1	Introduction	4
1.1	Background	4
1.1.1	The Global State of Water Pollution	4
1.1.2	The Importance of Public Participation in Mitigating Water Pollution	5
1.1.3	The Role of Social Media Integration in Mapping Water Pollution	6
1.2	Aims and Objectives	9
1.2.1	Research Questions	9
1.2.2	Research Objectives	9
1.2.3	Study Area	10
2	Literature Review	11
2.1	Water Pollution	11
2.2	Current Applications of Citizen Science	12
2.3	Integration of Social Media for Data Mining	14
2.4	Utilisation of Twitter for Data Mining	15
2.5	Data Creation and Analysis	16
2.6	Techniques for Topic Modelling	17
3	Methodology	19
3.1	Description of Data	20
3.2	Data Creation	21
3.2.1	Data Acquisition and Meta Data Collection	21
3.2.2	Geocoding and Spatial Contextualization	21
3.2.3	Data cleaning and Pre-processing	22
3.3	Text Mining Processes	22
3.3.1	Wordclouds	22
3.3.2	Sentiment Analysis	23
3.3.3	Topic Modelling - Implementing LDA and LSA in Python	24
3.4	Data Visualisation	24
4	Results and Discussion	25
4.1	Spatial Distribution of Tweets	25
4.2	Word clouds and Frequency of Words	27
4.3	Sentiment Analysis	33
4.3.1	Regional Sentiments by Location	35
4.4	Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA)	37
4.4.1	Analysis of Latent Dirichlet Allocation (LDA)	37
4.4.2	Latent Semantic Analysis (LSA)	39
5	Limitations and Future Research	42
5.1	Limitations and Challenges	42
5.2	Future Research	43
6	Conclusions and Recommendations	44
	References	46

List of Figures

1	Twitter users by country.	8
2	Twitter users by year	8
3	Geographic regions in the UK	10
4	Summary of the methodology adopted in the study	20
5	Raw format visualization of extracted location attributes prior to data pre-processing	21
6	Spatial Distribution of Water Pollution Tweets in the UK, between 18/06/2023 and 26/06/2023	25
7	Temporal Patterns of Water Pollution Tweets	26
8	Heatmap of Tweets in the UK	26
9	Word Cloud & Word Frequencies of Water Pollution Tweets in UK East Region . . .	27
10	Word Cloud & Word Frequencies of Water Pollution Tweets in UK East Midlands Region	28
11	Word Cloud & Word Frequencies of Water Pollution Tweets in UK West Midlands Region	28
12	Word Cloud & Word Frequencies of Water Pollution Tweets in UK North East Region	28
13	Word Cloud & Word Frequencies of Water Pollution Tweets in UK North West Region	29
14	Word Cloud & Word Frequencies of Water Pollution Tweets in UK South East Region	29
15	Word Cloud & Word Frequencies of Water Pollution Tweets in UK South West Region	30
16	Word Cloud & Word Frequencies of Water Pollution Tweets in London	30
17	Word Cloud & Word Frequencies of Water Pollution Tweets in UK Yorkshire and Humber Region	31
18	Word Cloud & Word Frequencies of Water Pollution Tweets in Scotland	31
19	Word Cloud & Word Frequencies of Water Pollution Tweets in Wales	32
20	Word Cloud & Word Frequencies of Water Pollution Tweets in Northern Ireland . . .	32
21	Distribution of Twitter Data Sentiment	33
22	Distribution of Twitter Data Sentiment:Pie Chart	34
23	Average Regional Sentiment scores in the UK	35
24	Word clouds for LDA model topics	37
25	LDA Bar plot showing weighting of words in each topic.	38
26	Topic Distribution of LDA model	39
27	Word clouds for LSA model topics	40
28	LSA Bar plot showing weighting of words in each topic.	40
29	Topic Distribution of LSA model	41

1 Introduction

1.1 Background

1.1.1 The Global State of Water Pollution

Water pollution is a widespread environmental concern of global scale, presenting significant challenges to the welfare of humanity, the integrity of ecosystems, and the accomplishment of sustainable development (Inyinbor Adejumo et al., 2018). This global phenomenon of water pollution is subject to the influence of various sources including urbanisation, inadequate waste management practices, and industrial operations such as mining. According to Törnqvist et al. (2011) and McGonigle et al. (2012), these activities result in the emission of pollutants, including heavy metals, pesticides, fertilisers, oil, and chemicals, into aquatic ecosystems, thereby causing a variety of adverse consequences.

Water pollution has far-reaching implications that impact nations at various stages of development, encompassing both high and low to middle income countries. In relation to human health, the presence of pollutants in water can give rise to health conditions, such as gastrointestinal infections, diseases characterised by diarrhoea, and infections. According to researchers worldwide, there is evidence suggesting that prolonged exposure to polluted water is associated with an elevated susceptibility to chronic illnesses such as cancer, endocrine disorders, and neurological impairments. Water pollution has a significant adverse impact on ecosystems as well. The discharge of pollutants has a significant impact on aquatic habitats, including rivers, lakes, and oceans (McGonigle et al., 2012; Vrain and Lovett, 2016). Eutrophication, resulting from elevated nutrient concentrations originating from agricultural runoff, can induce oxygen depletion and subsequent mortality of aquatic fauna, including fish and other organisms. The accumulation of toxic substances within the food chain has the potential to impact various organisms, including aquatic life as well as birds and mammals that rely on these ecosystems (Qadri et al., 2020; Bashir et al., 2020).

The implementation of international and national initiatives is imperative in addressing the issue of water pollution. Governments and regulatory bodies across the globe are enacting increasingly stringent regulations on various industries and wastewater treatment facilities with the aim of mitigating pollutant discharges (Jones and Gomes, 2013). In order to mitigate water pollution and foster a healthier environment for future generations, it is imperative to prioritise the promotion of sustainable agricultural practices, the implementation of improved waste management systems, and the elevation of public awareness regarding the significance of clean water (Qadri et al., 2020). Despite advancements made in achieving the objective of reducing the global population lacking sustainable access to safe drinking water and basic sanitation by 50 percent, as stipulated in the United Nations' Sustainable Development Goals (UN SDGs), the issue of water pollution continues to pose a substantial barrier (Fuller et al., 2022; M. Wang et al., 2022).

The imperative of addressing water quality and pollution, as well as mitigating greenhouse gases associated with water, constitutes pivotal measures in the pursuit of climate action aimed at attaining a future characterised by "net zero plus" outcomes. The primary objective of the "net zero plus" initiative is to attain a state of carbon neutrality, whereby the emissions produced are effectively balanced by the removal of greenhouse gases from the atmosphere (Mavrigiannaki et al., 2021). The primary objective is to achieve complete elimination of emissions, while concurrently engaging in activities that actively restore ecosystems and allocate resources towards sustainable practices that foster the overall regeneration of the planet (Levin et al., 2019). In order to mitigate the effects of climate change, it is

imperative to adopt an approach that involves intelligent sourcing and comprehensive, system-based resource management in the future (Fisher et al., 2020). Water pollution has the potential to disrupt the carbon cycle and consequently contribute to the emission of greenhouse gases. This underscores the importance of addressing water pollution as a significant factor in environmental management. Through the implementation of strategies aimed at preventing and mitigating water pollution, it becomes feasible to alleviate its adverse effects and progress towards a sustainable future characterised by net zero emissions (Gassie et al., 2016; Pattanayak et al., 2005).

1.1.2 The Importance of Public Participation in Mitigating Water Pollution

In addition to governmental initiatives, legislative actions, industry efforts, academic research, and citizen awareness plays a pivotal role in the mitigation of water pollution. The importance of public opinion and understanding in addressing water pollution cannot be overstated, as it encompasses multiple aspects. A comprehensive understanding of the origin and impacts of water pollution holds paramount importance for individuals, society, and industries. Acknowledging their role in both the cause and experience of water pollution cultivates a sense of accountability. The promotion of public comprehension about water utilisation and recycling fosters the adoption of more environmentally conscious behaviours, hence resulting in reduced resource consumption and pollution. Recognising the complex interplay between the natural environment and human beings engenders a sense of responsibility, prompting proactive measures for preservation (Otero et al., 2021).

To effectively promote substantial public engagement in the mitigation of water pollution, it is crucial to obtain a thorough comprehension of individuals' attitudes, perspectives, and level of awareness regarding this issue. The initial phase plays a crucial role in facilitating the modification of behaviour in individuals (Forleo and Romagnoli, 2021). Citizen science is an inclusive and cooperative methodology in which individuals from the general public actively engage in scientific research, as well as the collecting and analysis of data. Through the active involvement of citizens in the scientific process, this paradigm fosters inclusivity and democratisation, allowing individuals from all backgrounds to contribute their observations and knowledge towards the exploration and resolution of a wide range of scientific inquiries and difficulties (Stanton et al., 2022). The individuals involved in these endeavours, commonly known as "citizen scientists," encompass a diverse group of volunteers, students, and hobbyists that collaborate with experts in the scientific field to collect empirical information pertaining to various subjects, including but not limited to wildlife surveillance, climate dynamics, ecological variety, and even celestial exploration. Citizen science programs facilitate the acquisition of useful data from the public, which would otherwise be arduous or expensive for academics to gather independently. As a result, these initiatives enhance the breadth and magnitude of scientific investigations. This methodology facilitates the cultivation of public involvement in scientific endeavours, advocates for heightened consciousness of the environment, and enables individuals to actively contribute to the progression of knowledge and make well-informed choices that have positive implications for society at large (Huddart et al., 2016; Cook et al., 2021).

In recent times, there has been a noticeable increase in the dissemination of information pertaining to water pollution among the general public. The accomplishment of this objective has been facilitated through diverse methods, including the utilisation of reports, the organisation of events, the implementation of awareness campaigns, and the deployment of other informative strategies (Otero et al., 2021). However, social media has become a prominent platform that can be employed to educate the public about environmental risks and to collect their viewpoints on assessing water pollution and other environmental perils. The integration of social media can facilitate the mapping of water pollution by enabling efficient spatial identification, tracking, and consistent analysis of water pollution

events (Forleo and Romagnoli, 2021).

1.1.3 The Role of Social Media Integration in Mapping Water Pollution

The incorporation of social media platforms has revolutionised the manner in which individuals engage in communication and disseminate information. In addition to interpersonal relationships, social media has emerged as a potent instrument for addressing societal concerns and collecting Volunteered Geographic Information (VGI). Volunteered Geographic Information (VGI) pertains to the collection of geographical data and spatial information that is voluntarily given by individuals or communities (Plackett et al., 2020). This phenomenon signifies a significant shift in the collecting and dissemination of geographic data, wherein the utilisation of crowd-sourcing and citizen science is employed to enhance and expand upon conventional data sources (Zheng et al., 2017).

The growing popularity of the internet and the advent of web-based mapping tools have facilitated the active engagement of individuals in the generation of geographic content. Volunteers have the opportunity to contribute several sets of information, such as geographic coordinates, locations of interest, images, geotagged social media postings, and geographical annotations, through a multitude of internet platforms and mapping tools. VGI is distinguished by its grassroots methodology, which empowers individuals without professional expertise, commonly known as "citizen mappers," to engage in the generation and distribution of geographic data (See et al., 2017). The democratisation of geographical information facilitates a broader and more comprehensive portrayal of the world, encompassing a wider range of perspectives and insights that may otherwise be disregarded. VGI serves multiple purposes, encompassing mapping, navigation, disaster management, environmental monitoring, urban planning, and scientific research. VGI has demonstrated significant value, particularly in areas with restricted access to official data sources or in times of crises, where the availability of real-time, crowd-sourced information can greatly contribute to decision-making processes and reaction initiatives (Zheng et al., 2017; Tavra et al., 2021).

Social media has the potential to exert a substantial influence in the domain of water pollution mapping. Social media facilitates the effective identification, tracking, and analysis of water pollution incidents through its extensive coverage, immediate updates, and content generated by users (Barbier and Liu, 2011; Anderson et al., 2007). Social media platforms offer a wide-ranging user base that can provide significant contributions in terms of valuable information pertaining to incidents of water pollution. Concerned individuals have the ability to share their observations regarding polluted water bodies, such as contaminated rivers, lakes, or oceans, by employing hashtags, geolocation tags, and user-generated posts (Steiger et al., 2015; Otero et al., 2021). The data collected from various sources can be systematically observed and aggregated in order to generate a spatial representation of incidents and opinions related to water pollution. The utilisation of geotagging in posts enables prompt identification and precise localization of pollution-affected regions by authorities, researchers, and environmental organisations. This facilitates expedited response measures and focused remediation endeavours (McGonigle et al., 2012).

The real-time nature of social media enables the swift dissemination of information. When instances of water pollution transpire, individuals who are concerned have the ability to promptly disseminate visual imagery, videos, and written accounts. The real-time reporting feature facilitates the monitoring of pollution incidents in real-time, thereby offering significant data that can aid authorities in evaluating the magnitude and seriousness of the issue (O'Connor et al., 2010; McGregor, 2019). The utilisation of Twitter data and citizen science has demonstrated its significance as significant assets in the process of mapping and monitoring many natural phenomena across the world including forest fires, earthquakes, and other natural disasters.

Forest fires are frequently accompanied by the dissemination of real-time reports, photographs, and videos from impacted regions by Twitter users. This publicly accessible information is utilised by researchers and organisations to monitor the dissemination and magnitude of fires. During the occurrence of the California wildfires in 2016, social media data, specifically from Twitter, was utilised to generate interactive maps that depicted the spatial advancement of the fire, recommended escape paths, and identified vulnerable regions (Wang et al., 2016; Sachdeva and McCaffrey, 2018). This facilitated the prompt and efficient response of authorities, ensuring effective communication with the general population. At the same time, the impact of earthquakes has been expeditiously evaluated by the utilisation of Twitter data. Following a seismic event, individuals frequently utilise social media platforms to share their personal encounters, provide accounts of the perceived severity, and disclose the geographical location associated with the occurrence. Through the examination of these tweets, researchers have the ability to generate "Did You Feel It?" maps, which facilitate the estimation of the magnitude of shaking in the places that have been impacted. The utilisation of this data significantly improves the accuracy of seismic hazard assessments and provides valuable support for the development of emergency response plans (Mendoza et al., 2019).

The utilisation of Twitter data has been employed as a means to monitor and evaluate the effects of hurricanes and storms. Individuals actively communicate and disseminate up-to-the-minute information regarding occurrences of flooding, wind-related destruction, and electrical power disruptions. The data is analysed by researchers and authorities in order to identify places that require immediate assistance, distribute resources, and assess the impact of the storm on populations (Yum, 2020). Furthermore, flooding incidents are frequently reported by residents on social media platforms such as Twitter during periods of intense rainfall and flooding. The utilisation of crowd-sourced information contributes to the development of flood maps, the identification of areas susceptible to flooding, and the improvement of techniques for managing floods (Wang et al., 2018).

The augmentation of analytical capabilities can be achieved by integrating social media data with pre-existing water quality monitoring systems by authorities. User-generated content, comprising of various forms of media such as photographs and videos, has the potential to complement conventional monitoring techniques by providing supplementary visual documentation of pollution occurrences (Alvarez-Melis and Saveski, 2016). In addition, the utilisation of sentiment analysis and data mining methodologies can be employed to analyse social media posts with the aim of detecting patterns, trends, and potential contributors to environmental pollution. The consistent analysis of water pollution can assist authorities in comprehending the underlying causes and consequences of this issue, thereby facilitating informed policy-making and the identification of specific areas that require targeted intervention (Fuller et al., 2022; O'Connor et al., 2010; Gassie et al., 2016).

Among the diverse social media platforms, Twitter has gained significant prominence boasting a substantial user base of more than 330 million active individuals (Karami et al., 2020). Figure 1 illustrates a graphical summary of twitter users across the world.

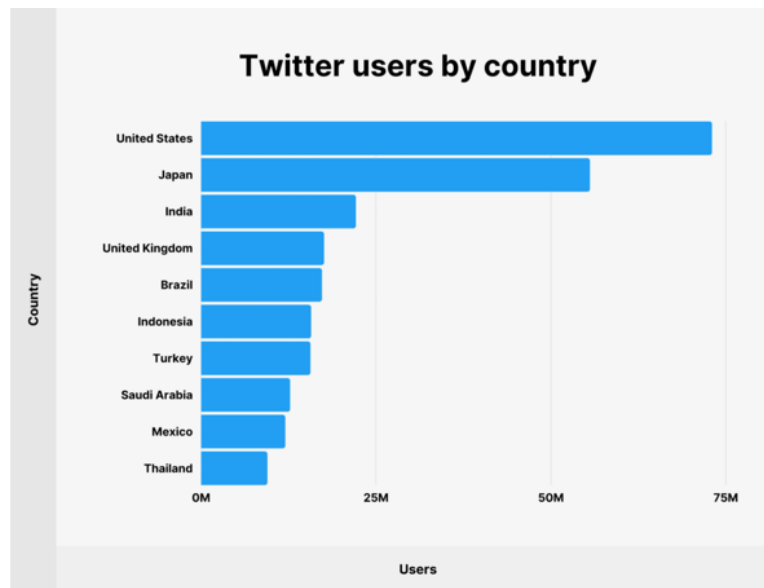


Figure 1. Twitter users by country.

Consequently, it has become one of the most extensively utilised platforms among the various options available. Twitter is acknowledged as a microblogging platform that enables the dissemination of user-generated content through tweets, encompassing a daily volume exceeding 500 million posts and comments (Cheng et al., 2021; Steiger et al., 2015). Figure 2 indicates the rapid twitter user growth from 2010 to 2021.

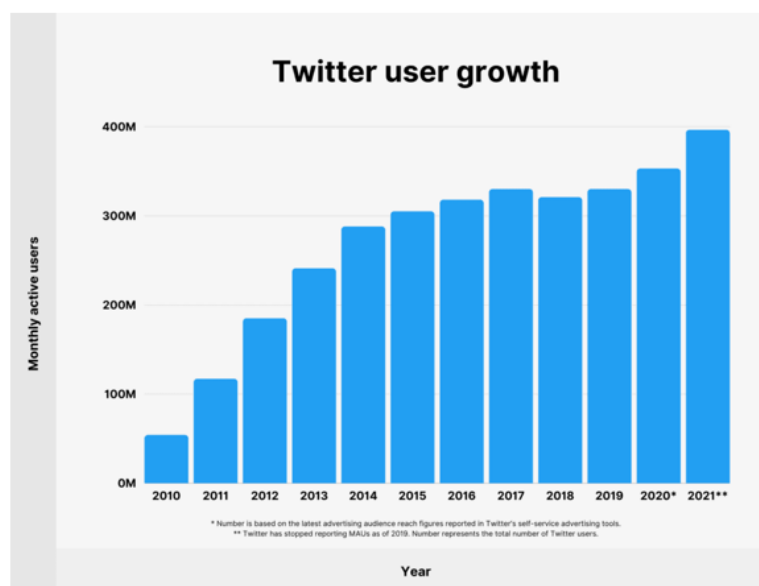


Figure 2. Twitter users by year

The utilisation of this platform has demonstrated its significance in facilitating users' ability to articulate their viewpoints and ideas, thus functioning as a valuable tool for cultivating public awareness on a wide range of subjects (Mislove et al., 2011). Additionally, it has been argued that Twitter has the potential to effectively facilitate the transformation of public opinions and thoughts into Volunteered Geographic Information (VGI) (Karami et al., 2020).

Researchers globally have observed that there is potential in utilising Twitter data mining for spatial analysis to assess water pollution levels in the United Kingdom (Otero et al., 2021). This observation is based on empirical evidence, which highlights that during the year 2021, approximately 19.05 million individuals in the United Kingdom were found to possess a Twitter account (Cheng et al., 2021). By conducting a systematic analysis of water pollution patterns, trends, events, and opinions, it is possible to identify regions with high levels of pollution and improve our understanding of resource management. The acquisition of this knowledge possesses the capacity to directly impact the process of decision-making and provide guidance for future research initiatives focused on achieving a net-zero-plus future and mitigating greenhouse gas emissions in alignment with the objectives outlined by the United Nations Millennium Development Goals (Liu et al., 2014). The above statement underscores the intrinsic potential of social media as a tool for monitoring the environment, underscoring the importance of integrating public viewpoints into sustainable development strategies (Otero et al., 2021; Spielhofer et al., 2016).

1.2 Aims and Objectives

Based on the aforementioned background study, the present study aims to utilise Twitter data to examine the spatial-temporal distribution of water pollution incidents in the UK, evaluate regional and stakeholder disparities in perceptions and investigate the feasibility of cooperation in water quality management.

The study is centred around the following research questions and objectives.

1.2.1 Research Questions

1. How do perceptions of water pollution vary across different regions and stakeholders in the UK?
2. How can Twitter data be utilised to identify and validate secondary geospatial data sources for assessing water pollution?
3. What are the potential benefits and limitations of using Twitter data for assessing the water pollution in UK?

1.2.2 Research Objectives

1. To collect and pre- process geolocated Twitter data related to water pollution within the UK.
2. To conduct a spatio-temporal analysis to identify regional hotspots of water pollution and causes based on the Twitter data.
3. To conduct topic modelling and sentiment analysis to identify the regional differences in public perception and different stakeholder perception on water pollution.
4. To identify and access other secondary geospatial data sources via Twitter

1.2.3 Study Area

This research is specifically confined to the geographical boundaries of the UK. Its main purpose is to investigate the viability of utilizing data mining and spatial analysis techniques on Twitter data to assess the extent of water pollution within the UK's geographical limits. The central emphasis of this study revolves around exploring concerns and route to identify new incidents related to water quality and instances of pollution within the context of the UK. The ultimate goal is to attain a profound and exhaustive comprehension of these issues within this precise geographical area.

UK, despite its status as a high income country, also faces significant challenges due to water pollution. Notwithstanding substantial endeavours to enhance water quality, the issue of pollution stemming from agricultural activities, urban runoff, and industrial discharges persists as a matter of concern (McGonigle et al., 2012). The monitoring and regulatory efforts of the Environment Agency in the UK are aimed at overseeing water bodies and implementing strategies to mitigate pollution. However, despite these measures, specific regions continue to encounter challenges such as elevated nutrient concentrations, the presence of microplastics, and contamination resulting from past industrial activities (Cooper and Findlater, 2013).

By adopting a localised approach, this research ensures a concentrated and all-encompassing scrutiny of patterns, emotions, and spatial dispersions linked to pollution. This methodology guarantees that the outcomes and implications are tailor-fitted to the distinct environmental and social dynamics distinctive to the UK. The figure 3 illustrates both the geographic borders of the UK and the administrative divisions employed for the analytical procedures carried out in this research.



Figure 3. Geographic regions in the UK

2 Literature Review

The significance of this research is underscored by the literature review, which presents an intricate examination of the prevailing body of knowledge concerning the subject. Within this section, we encapsulate the pivotal works that have undergone scrutiny throughout this study, focusing on the global employment of social media data by researchers from diverse disciplines to amass public sentiment.

2.1 Water Pollution

Water pollution constitutes a pervasive challenge affecting countries at every stage of development, encompassing both advanced and emerging economies. The surge in urbanization and industrialization in diverse regions has notably exacerbated this issue, triggering profound ramifications for human well-being, aquatic habitats, and biodiversity (Törnqvist et al., 2011; McGonigle et al., 2012). These effects reverberate throughout ecosystems, resulting in the decline of numerous species (Bashir et al., 2020). Nutrient pollution, stemming from agricultural activities, contributes to the proliferation of harmful algal blooms. These blooms facilitate oxygen depletion and the emergence of oxygen-deficient aquatic zones, commonly referred to as "dead zones". Consequently, the repercussions of water pollution extend beyond direct human impact to encompass significant ecological disruption. The existence of hazardous elements, such as heavy metals and chemical compounds, further compounds these hazards, endangering both human communities and surrounding ecosystems (Bashir et al., 2020; Inyinbor Adejumo et al., 2018; McGonigle et al., 2012).

The pollution of marine habitats by plastics is a growing concern due to the potential threats it poses to both human and environmental health (Periathamby et al., 2019). The magnitude of plastic pollution has been exacerbated by the extensive utilisation of the material, insufficient and hazardous disposal practises, as well as its capacity to disperse and remain in the environment (Edelson et al., 2021). The increasing apprehension among nations has led to efforts aimed at minimising and eliminating plastic waste from the natural environment. Interventions encompass a range of measures, such as the implementation of prohibitions on the usage of single-use plastics, the establishment of container deposit schemes, and the allocation of resources towards the enhancement of waste management systems (Periathamby et al., 2019).

The mitigation of water pollution necessitates a comprehensive strategy that incorporates the implementation of policy reforms, advancements in technology, and the cultivation of public consciousness. Over the course of time, substantial research endeavours have been dedicated to the amelioration of water pollution and the advancement of sustainable water management methodologies (McGonigle et al., 2012). Researchers have been actively engaged in the exploration and development of innovative water treatment technologies with the aim of enhancing the efficiency of pollutant removal. Various methods such as membrane filtration, reverse osmosis, and ultraviolet disinfection have demonstrated potential in the treatment of wastewater originating from industrial and municipal sources (Fuller et al., 2022).

The collaboration between scientists and policymakers is crucial in the establishment and implementation of stringent environmental policies and standards. Governments can establish regulatory measures to enforce restrictions on the presence of pollutants in water bodies, so ensuring that industries and individuals are held responsible for their respective roles in contributing to water pollution. The implementation of appropriate waste management practises, the adoption of environmentally conscious agriculture techniques, and the promotion of sustainable industrial processes are crucial in mitigating pollution levels (Dhanwal et al., 2017; Sen et al., 2013).

International collaboration is essential in addressing the issue of water pollution. Academic researchers and institutions engage in cross-border collaborations to facilitate the exchange of knowledge, experiences, and best practises, thereby contributing to the development of more complete and efficacious solutions (Brack et al., 2019).

The promotion of public awareness and education plays a crucial role in cultivating a sense of responsibility among communities by disseminating information regarding the detrimental effects of water pollution and the significance of water conservation. There is a higher likelihood of educated individuals demonstrating support for and engaging in initiatives aimed at safeguarding their nearby water resources Otero et al. (2021).

Communities globally encounter a substantial challenge in the form of water contamination. The aforementioned phenomenon poses a risk not only to human well-being, but also to the intricate equilibrium of aquatic ecosystems. Researchers and organisations are actively engaged in addressing the significant problem of water contamination, demonstrating a deep understanding of its seriousness and dedicating their efforts towards finding effective solutions (McGonigle et al., 2012). By leveraging cutting-edge technologies, implementing enhanced policies, and fostering heightened public awareness, there exists a promising prospect of effectively tackling water pollution and ensuring the preservation of our invaluable water resources for future generations. Nonetheless, the achievement of significant transformation and the establishment of a water future that is cleaner, healthier, and more sustainable require persistent dedication and involvement by governmental bodies, enterprises, and individuals (Cooper and Findlater, 2013; Törnqvist et al., 2011)..

2.2 Current Applications of Citizen Science

Citizen science is an inclusive and cooperative methodology that engages non-expert individuals, such as members of the public or volunteers, in the undertaking of scientific research endeavours. The utilisation of this participatory approach has garnered substantial traction on a global scale, effectively leveraging the potential of ordinary individuals to make valuable contributions to various scientific pursuits (Bonney et al., 2016).

The utilisation of citizen science has emerged as a potent mechanism for the surveillance of biodiversity and environmental shifts. Many web based platforms allow users the opportunity to record and disseminate observations pertaining to flora, fauna, and avian species, thereby furnishing researchers and conservationists with invaluable data. Researchers worldwide emphasize that these contributions have facilitated the identification of novel species, monitoring the geographical range of species, and comprehending the ramifications of climate change on ecosystems (Theobald et al., 2015; Chandler et al., 2017). Climate scientists are progressively involving civilians in the collection of meteorological data, documentation of extreme events, and measurement of climate pattern alterations. The research have facilitated the collection of comprehensive datasets, so upgrading climate models and advancing our comprehension of climate variability (Reyes-Garcia et al., 2020).

The field of health and medicine encompasses several aspects related to the promotion, maintenance, and restoration of physical and mental well-being. It involves the study and application of scientific knowledge Citizen science has demonstrated significant potential within the field of health and medicine. Many initiatives leverage the collective computational resources contributed by volunteers to investigate the intricate process of protein folding, thereby facilitating advancements in drug development and enhancing our comprehension of ailments such as Alzheimer's disease and cancer. Additional citizen science activities in the realm of health encompass the monitoring of disease outbreaks, the assessment of air quality, and the facilitation of public health research (Wiggins and Wilbanks,

2019; King et al., 2019).

In the field of archaeology which encompasses the study of material remains and artefacts from past human societies, with the aim of understanding and interpreting their cultural heritage, the participation of citizen scientists is evident through their documentation of historical places, artefacts, and fossils. One such instance is the Ancient Lives project, when volunteers engage in the transcription of papyri originating from ancient Egypt. These endeavours contribute to the conservation of cultural assets and offer valuable perspectives on historical contexts (Wessman et al., 2019).

The urban planning and transportation discipline encompasses the study and practise of designing and managing the built environment in order to facilitate efficient and sustainable movement of people and goods within urban areas. The utilisation of citizen science has been employed to gather data pertaining to urban infrastructure, traffic patterns, and the utilisation of public transportation. The aforementioned data holds significant value in the context of urban planning, as it enables the optimisation of transport networks and contributes to the overall sustainability and accessibility of cities (Franco and Cappa, 2021).

The utilisation of citizen science has become a prominent and cooperative strategy within the water pollution research. Citizen science facilitates the acquisition of water quality data including the parameters such as temperature, pH levels, nutrition levels and levels of contaminants present in water such as heavy metals and micro plastics with generic and active involvement of the general population (Otero et al., 2021). Furthermore in the fields of social sciences and education, citizen science is employed to facilitate the conduction of surveys, collection of opinions, and examination of human behaviour. The research domains encompass political science, psychology, sociology, and education, wherein the utilisation of extensive data collection methods can contribute to the development of more substantial and reliable conclusions as expressed by Albert et al. (2021). The practice of citizen science promotes active involvement of the public and enhances their understanding of scientific concepts.

Citizen science also engages the public in the monitoring and mitigation of environmental degradation. Numerous research studies have demonstrated the efficacy of initiatives led by citizens in the collection of vital data pertaining to water quality and levels of pollution. By employing affordable monitoring devices and user-friendly applications, consumers are able to provide real-time data that supplements conventional scientific methodologies (Brockhage et al., 2022). This methodology not only broadens the extent of data acquisition but also enables local people to assume responsibility for their water resources. These projects have exhibited their efficacy in the identification of pollution sources, the detection of patterns, and the provision of information for policy-making processes. The involvement of individuals in the identification of pollution hotspots not only assists in locating these areas but also enhances public consciousness regarding the pressing requirement for sustainable water management. Citizen science serves as a collaborative mechanism that facilitates the connection between scientific research, public perception, and practical measures aimed at addressing water pollution (Gacutan et al., 2023).

Thus, based on the available literature, it is evident that citizen science possesses significant potential in the future. The integration of artificial intelligence and machine learning has the potential to significantly accelerate data analysis and validation processes, hence facilitating the execution of more ambitious projects (Nunavath and Goodwin, 2018). The proliferation of mobile technology and internet connectivity is anticipated to facilitate enhanced engagement of citizens in citizen scientific programmes, hence fostering worldwide collaboration. The utilisation of citizen research has become a significant catalyst for scientific exploration, environmental surveillance, and community involvement. The applications of this technology encompass a wide array of disciplines. As scholars

persist in enhancing approaches and confronting obstacles, it is undeniable that citizen science will assume a progressively influential function in moulding the trajectory of scientific inquiry and societal advancement (Otero et al., 2021; Christian et al., 2012; Bonney et al., 2016; Franco and Cappa, 2021).

2.3 Integration of Social Media for Data Mining

The practice of data mining in social media is of utmost importance as it enables the extraction of significant insights and patterns from the extensive volume of data present across many social media platforms. In order to obtain results that hold significance, it is imperative to take into account a number of crucial aspects. Different social media platforms and data mining objectives may necessitate distinct methodologies and algorithms. Various datasets and research inquiries necessitate the use of distinct analytical tools. For instance, classification techniques are suitable for structured data, while clustering methods are effective in detecting underlying trends and patterns. A comprehensive grasp of both the data and the mining tools at one's disposal is crucial for the attainment of successful analysis outcomes (Gacutan et al., 2023).

The process of preprocessing is of the highest priority, encompassing various factors such as the implementation of anonymisation techniques and the safeguarding of privacy. While it is true that social media data is accessible to the public, it is crucial to uphold the rights of individuals and respect the copyrights of the respective social media platforms. Furthermore, it is imperative to consider the consequences of spam and the temporal representation (Brockhage et al., 2022; Injadat et al., 2016). The element of time is of greatest significance, as outcomes might exhibit substantial variations over distinct temporal intervals. Temporal shifts have a significant impact on community detection, group behaviour, and marketing. The definition of a community or the interests associated with it may undergo changes throughout time. Therefore, conducting a temporal analysis of data can provide valuable insights.

Social networking sites and blogs represent two distinct categories of social media data that offer dynamic and abundant sources of information. Social networking services, such as Facebook and LinkedIn, are comprised of interconnected members who possess distinct profiles. Data mining can be utilised for the purpose of identifying groups, creating profiles, and developing recommendation systems in order to get insights about collaborative dynamics and group formations. The consideration of privacy concerns should always be prioritised (Dittmann et al., 2023).

The utilisation of data mining techniques in the realm of social media presents a potent instrument for acquiring significant and valuable insights. In order to obtain meaningful results, it is crucial to possess a comprehensive understanding of the data, make informed decisions regarding the selection of an appropriate mining strategy, and carefully examine various issues such as time constraints and privacy concerns. Social networking sites and blogs offer valuable sources of data that can be analysed to get insights into group dynamics, individual interests, and interpersonal relationships. This has significant implications for both the scientific community and commercial enterprises (Gundecha and Liu, 2012).

Volunteered Geographic Information (VGI) encompasses data that is willingly supplied by individuals and holds significant value across several disciplines, such as disaster management, environmental management, pollution mitigation and crisis response. The process of geocoding, which involves linking data with location information, is understood as of paramount importance in rendering this data suitable for geospatial analysis. Nevertheless, several social media platforms place a higher emphasis on maintaining user anonymity, resulting in the absence of geographic markers in posts.

The process of geocoding these posts is dependent on the examination of textual information (Tavra et al., 2021).

2.4 Utilisation of Twitter for Data Mining

Social media platforms, with a particular emphasis on Twitter, is a noteworthy function in disseminating information and fostering public involvement in matters pertaining to the environment (Otero et al., 2021). This study undertakes an analysis of a data set comprising more than 140,000 tweets in order to investigate the level of public interest and awareness pertaining to marine trash. The study employs text mining techniques, sentiment analysis, and geo-tagging to identify prevailing subjects, trends in sentiment, and areas of concentrated debate. The examination of images related with tweets is conducted to evaluate their capacity for identifying and categorising litter quantities and varieties, with a specific focus on coastal regions. The results derived from this study possess the potential to assist governmental organisations, academic institutions, and non-governmental organisations (NGOs) engaged in the preservation of maritime environments in enhancing their communication strategies on the social media platform, Twitter. By comprehending the individuals who engage in tweeting about marine litter and their methods of doing so, these institutions can develop communication strategies that are more efficient in reinforcing involvement, improving public comprehension of potential remedies, and facilitating collaborative efforts to combat marine pollution.

Twitter is a beneficial data resource in scientific research pertaining to the geospatial field. It can be used effectively for the purpose of geographic analysis. Additionally, the data mined from Twitter can be categorized based on the information they offer, which can be utilised for diverse data analyses. Twitter enables the aggregation of information received from a large number of individuals, thereby offering timely updates that were previously unattainable through conventional means of data collection such as surveys and censuses. This holds significant relevance for governmental bodies and institutions aiming to enhance societal well-being and effectively tackle specific challenges (Saldana-Perez et al., 2019).

Twitter can be applied to examine the public attitude and interest in environmental incidents. The Toledo water ban is utilised as a case study in this research. This observation underscores the persistent occurrence of eutrophication problems induced by human activities in lakes and coastal ecosystems, hence resulting in such occurrences. The analysis of Twitter data provides insights on contemporaneous terms and public sentiment during the Toledo crisis, whilst Google Trends offers a more extended perspective by showcasing sustained attention towards the issue of algal bloom, with a notable increase in interest during the occurrence (Cha and Stow, 2015). The integration of data from Twitter and Google Trends yields significant insights. The utilisation of online data expands the range of information that is available to scientists and policymakers, fostering engagement among varied groups and facilitating connections between them. In contrast to traditional routes, online data mining offers a more cost-effective and efficient approach, while also encompassing a broader spectrum of ideas. The examination of dynamic internet information plays a crucial role in providing insights for ecosystem management and resource allocation in investigations and monitoring initiatives (O'Connor et al., 2010).

The emergence of micro-blogging platforms, such as Twitter, has marked a significant avenue for fostering public engagement with various environmental concerns. Through the extraction of valuable insights, these platforms play a pivotal role in facilitating the dissemination of information on diverse environmental issues. However, within the realm of water pollution analysis, there exists a notable research gap, where the utilization of Twitter data for this specific purpose remains relatively unexplored. This scarcity of dedicated research underscores the significance of the current study, as

it aims to address this gap by delving into the unique potential of Twitter data to unravel insights into water pollution dynamics.

2.5 Data Creation and Analysis

Researchers worldwide have harnessed various techniques, tools, and methodologies to convert social media data into valuable insights. Artificial Intelligence (AI), notably Machine Learning (ML), plays a pivotal role in the analysis of extensive social media data streams, aiding in addressing environmental concerns and evaluating natural disaster events (Nunavath and Goodwin, 2018). There is a growing interest among individuals to actively engage in social media and believe that their contributions can be beneficial to others. Posts on social media platforms contain both metadata and text, which can be utilised for georeferencing purposes (Saldana-Perez et al., 2019). However, it is worth noting that only a limited number of posts have location information. The study categorizes the posts into three types: self-centric, social, and collective information. Self-centric posts reveal the emotions and activities of the users, while social posts disseminate information about significant events such as sociopolitical movements. On the other hand, collective information posts provide valuable data about natural disasters and urban situations. Researchers in the social sciences primarily focus on analyzing self-centric and social posts to gain insights into people's emotions, activities, and opinions (Periathamby et al., 2019).

Similarly, social and collective information posts are valuable for forecasting and modeling urban factors and natural phenomena, considering their geographical context and timestamps. The research underscores the importance of combining computing and geographic sciences to develop effective data management procedures for extracting and analyzing posts for scientific purposes. Different APIs, text mining, and NLP techniques have been developed to handle the textual aspect of posts, and methods for georeferencing posts lacking location coordinates have been successfully designed and implemented (Alvarez-Melis and Saveski, 2016). To enhance accuracy and ensure comprehensive data analysis, information from this social media platform can be cross-referenced with authoritative sources and other credible information. It is widely acknowledged as a valuable source for monitoring real-world dynamics from the perspectives of its users, providing invaluable first-hand information (Albert et al., 2021).

Data quality on text classification outcomes, emphasises the preference for balanced datasets. While larger datasets generally improve performance, smaller collection windows also show positive correlations. Domain-specific and API-driven sources contribute significantly to datasets, with some researchers utilising previously published data for benchmarking. Normalization focuses on handling nonnatural language entities in social media, including usernames, URIs, and hashtags. Converting social media content to plain text aids algorithm processing (Rogers et al., 2021). Classic NLP tasks like tokenisation, POS tagging, lemmatisation, and stemming are commonly employed, with a focus on lexico-semantic sentiment and emotion analysis. The study employs SVMs, Bayesian classifiers, and decision trees as the main algorithms, with SVMs generally outperforming others. Neural networks appear more frequently in recent studies, facilitated by user-friendly software packages. The integration of text classification tools into large analytic platforms gains importance, driven by the widespread adoption of machine learning beyond computer science. Interdisciplinary research blends qualitative and quantitative analyses. The future of social media text mining relies on multilingual and platform-agnostic solutions, supported by a diverse range of gold-standard corpora (Liu et al., 2020).

An innovative method known as Geographical Knowledge Discovery (GKD) combines an analysis of social sources, digital newspapers, and official databases to extract citizens' viewpoints on specific

social issues. By integrating the geographical and temporal dimensions, it not only uncovers data correlations but also reveals additional insights, which boosts confidence in certain publications. The GKD process effectively selects and links relevant semantic web datasets and spatial datasets related to social problems. Future endeavors will focus on enhancing the classification process through the application of advanced deep learning techniques and incorporating diverse data sources. Furthermore, the potential benefits of utilizing ontology to improve the data classification process will also be investigated (Zagal et al., 2016).

The k-means technique proves effective in mapping regions afflicted by water pollution in Indonesia based on province divisions (Nunavath and Goodwin, 2018). The researchers successfully distinguished four provinces, namely North Sumatra, West Java, Central Java, and East Java, in the high cluster (K1). Additionally, they assigned thirty provinces to the lower cluster (K2). The validity of the clustering was assessed using the Davies-Bouldin Index (DBI), yielding a value of 0.328. This DBI score suggests that the quality of the resulting clusters is commendable. This research could have significant implications for policymakers and environmental authorities in targeting specific areas for pollution control and remediation initiatives.

2.6 Techniques for Topic Modelling

Topic modelling is a widely utilised unsupervised machine learning methodology employed for the purpose of uncovering hidden semantic structures within extensive sets of unorganised textual data (Alvarez-Melis and Saveski, 2016). The objective of this approach is to extract significant topics from documents and classify them according to their patterns of co-occurrence. Numerous techniques for topic modelling have been put forth, including Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Singular Value Decomposition (SVD). This section offers a thorough examination of various subject model possibilities, analysing their respective merits, limitations, and domains of use according to the existing literature (Liu et al., 2020).

To successfully achieve the objectives of the present study, it is important to possess a comprehensive understanding of the topic modelling techniques including Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Singular Value Decomposition (SVD). These methodologies facilitate the extraction of significant patterns, subjects, and feelings from geographically tagged Twitter data, so assisting in the identification of areas with high levels of water pollution, the determination of underlying causes, and the exploration of varied regional perspectives. The utilisation of LDA, LSA, and SVD techniques enables researchers to reveal latent patterns and knowledge embedded in the data, hence facilitating the efficient study of spatial and temporal aspects (Wang and Zhu, 2017). Moreover, these methods aid in identifying diverse perspectives held by the general public and different stakeholders concerning the dynamics of water pollution in the United Kingdom.

The Singular Value Decomposition (SVD) is a mathematical technique used to decompose a matrix into three separate matrices, namely the left singular vectors, the singular values, and the right singular vectors. SVD is a matrix factorization technique that is commonly used in topic modelling (Sorour et al., 2014). Its purpose is to decrease the SVD is extensively employed for diverse applications, such as Latent Semantic Analysis (LSA), it is not as prevalent in the field of topic modelling as Latent Dirichlet Allocation (LDA). This is mostly attributed to the inherent limits of SVD in terms of its capacity to provide interpretable topics. The process of SVD results in a reduction of dimensions without taking into account the inherent semantics of the data, perhaps resulting in the generation of subjects that are less meaningful (Wang and Zhu, 2017). SVD is utilised in various disciplines, including collaborative filtering, picture compression, and data visualisation. Collaborative filtering-based recommender systems employ SVD to estimate user preferences for items, hence facilitating the

provision of personalised recommendations (Kousika and Premalatha, 2021). The application of SVD in image compression enables the reduction of storage space needed for image storage by expressing images in a lower-dimensional space. Moreover, the SVD technique aids in the visualisation of data by converting data with a high number of dimensions into a lower-dimensional representation. This transformation facilitates the study and interpretation of the data (Xu et al., 2006; Wang and Zhu, 2017).

Latent Semantic Analysis (LSA) is a computational method used in Natural Language Processing (NLP) and information retrieval to analyse the relationships between words and documents (Sorour et al., 2014). LSA is a well-established topic model that use the technique of Singular Value Decomposition (SVD) to uncover latent patterns inside a matrix representing the relationship between terms and documents. The process involves mapping the documents and phrases onto a lower-dimensional space in order to identify similarities among them. LSA demonstrates strong performance in reducing dimensionality and eliminating noise. However, it exhibits limitations in effectively capturing the contextual meaning of words. The potential inefficiency in handling polysemy, wherein a term has numerous meanings, may result in issues that are less easily interpreted. LSA has been widely utilised in diverse fields, encompassing information retrieval, document clustering, and text categorization. The utilisation of this technique in the domain of information retrieval has been extensively employed due to its capacity to enhance document ranking and retrieval within search engines. The utilisation of latent semantic structures in LSA enables document clustering to effectively group together documents that share similar characteristics. Moreover, LSA can facilitate text classification endeavours by effectively classifying articles into pre-established groups or topics (Wang and Zhu, 2017; Cvitanic et al., 2016).

Latent Dirichlet Allocation (LDA) is a probabilistic generative model used for topic modelling. LDA is a commonly utilised probabilistic generative model employed in the field of topic modelling. LDA model posits that every text is composed of a combination of many subjects, and each word within the document is generated from a particular topic's probability distribution. The proposed method demonstrates superior interpretability of subjects in comparison to LSA and effectively manages the challenge of polysemy. LDA is widely acknowledged as a very stable and scalable technique for topic modelling, rendering it particularly well-suited for the analysis of extensive datasets (Lin et al., 2016). LDA is widely utilised across various domains, including but not limited to social media analysis, market research, healthcare, and recommender systems. In the field of social media analysis, the LDA algorithm can be utilised to examine user-generated content on various platforms such as Twitter, Facebook, and online forums. This application aims to identify emerging subjects of discussion and assess the prevailing sentiments associated with them (Tong and Zhang, 2016). The utilisation of LDA in market research facilitates the identification of patterns within customer feedback, product reviews, and survey replies, hence enhancing comprehension of consumer preferences and behaviour. LDA has been employed in the healthcare field to do medical text analysis, wherein it is utilised to identify topics from electronic health records or medical literature. This application of LDA aims to support clinical decision-making processes. Furthermore, LDA holds significant value in the context of recommendation systems due to its ability to generate content-based recommendation systems. This is achieved by effectively recognising latent topics within user preferences and item descriptions (Tong and Zhang, 2016; Lin et al., 2016; Kalepalli et al., 2020).

Numerous studies have been undertaken globally utilising Twitter data to collect public sentiment on various subjects, as discussed in the preceding sections. The table presented in table 1 provides an overview of various methodologies and procedures employed in prior research. The inclusion of any procedures or methods in the specific study is indicated by the columns marked in grey.

Used Technique						Reference
Number of tweets	Topic Modeling	Sentiment Analysis	Temporal Analysis	Geospatial Analysis	Network Modeling	
7,881	(Hawkes Binomial Topic Model)					Sha et al. (2020)
23,830,322	(LDA)					Ordun et al. (2020)
1,150,825						Abd-Alrazaq et al. (2020)
203,756	(LDA)					Wicke and Bolognesi (2020)
1,873,022	(LDA)					Duong et al. (2020)
126,049	(LDA)					Medford et al. (2020)
100,000,000						Kabir and Madria (2020)
6,468,526						Lopez et al. (2020)
2,792,513	(Identifying frequently appeared words)					Singh et al. (2020)
1,187,482	(Word Embedding)					Cinelli et al. (2020)
8,919,411	(LDA)					Chen et al. (2020)
30,800,000	(Word Embedding)					Sharma et al. (2020)

Table 1. Different topic modelling, spatial and sentiment analysis techniques Incorporated in literature

3 Methodology

This section provides a detailed description of the methodology adopted in this study. This research seeks to utilise data mining and spatial analysis methodologies to evaluate water pollution in the UK through the examination of data derived from Twitter. In order to achieve this objective, the research employs the Twitter API and the Python programming language to obtain pertinent tweets. The data obtained was subsequently subjected to a data generation process, which involves data preprocessing and geocoding, in order to improve its quality and incorporate spatial context. Following this, the utilisation of topic modelling and sentiment analysis is employed to extract valuable insights regarding the talks pertaining to water pollution on the social media platform, Twitter. Furthermore, the study investigates the utilisation of image analysis techniques in order to determine the visual components associated with water pollution. The incorporation of these methodologies facilitates a thorough comprehension of water pollution hotspots, patterns or trends of pollution and the perception of the public within the UK. A graphical illustration of the methodology adopted in the study is presented in figure 4.

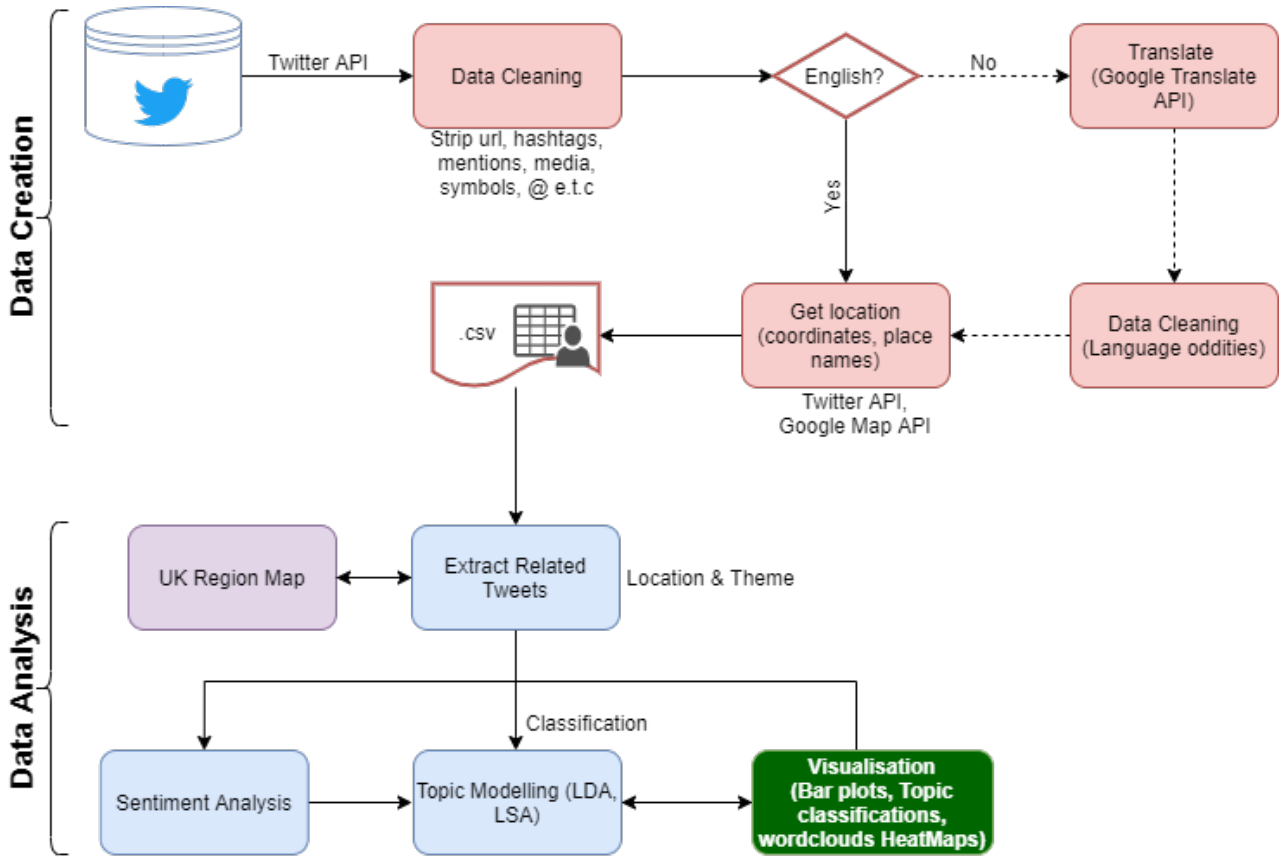


Figure 4. Summary of the methodology adopted in the study

3.1 Description of Data

The data for this study was obtained using the Search API of Twitter, with the assistance of the tweepy library in Python, along with other Python libraries. The purpose was to collect original tweets that are pertinent to the topic of water pollution. Multiple queries are constructed with the purpose of retrieving tweets that include particular keywords such as 'pollution', 'water', and 'waste' in order to guarantee the relevance of the collected data.

Throughout the duration of this research, it is crucial to emphasize a significant constraint - the lack of historical data that is typically accessible through the Twitter Academic Research Access Tier. This shortage of data can primarily be attributed to the dynamic and ever-changing nature of the Twitter API, currently undergoing modifications due to changes in ownership. As a result, our research was constrained to rely exclusively on a dataset encompassing only one week's worth of information, consisting of approximately 11,473 data points. This dataset specifically covers the period from January 18, 2023, to January 26, 2023. This time frame corresponds to the data available through the basic access tier during the specified period. Despite this limitation, the research proceeded to conduct a comprehensive analysis within these constraints in order to extract valuable insights for this study.

The raw structure of the location data are illustrated in figure 5. In the absence of any data cleansing procedures, the location data exhibits mirrored and displaced coordinates on the map, appearing in various positions.

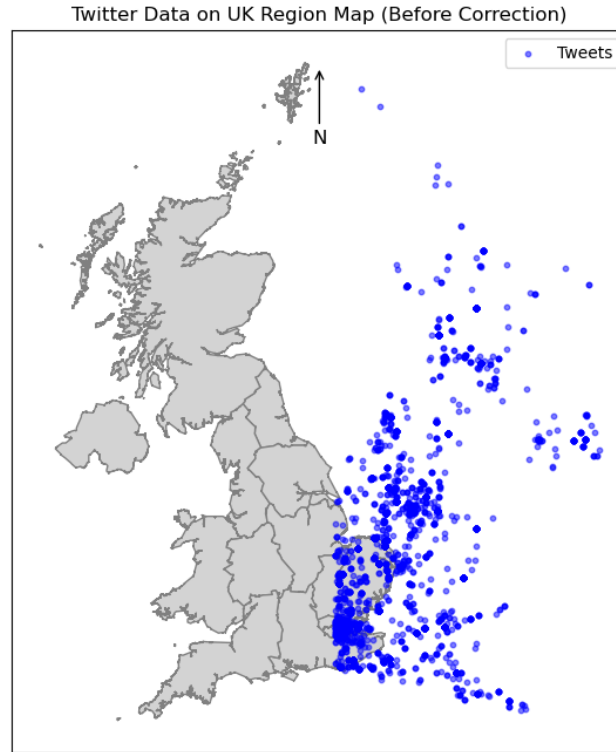


Figure 5. Raw format visualization of extracted location attributes prior to data pre-processing

3.2 Data Creation

The present study employs a technique encompassing several key steps to facilitate the creation and pre-processing of data, with the objective of examining geospatial patterns of tweets within the UK. The process involves several essential stages, namely data acquisition, metadata collecting, geocoding, translation, and data cleaning. Each of these procedures plays a crucial role in enhancing the quality and accuracy of the dataset, hence facilitating more precise and meaningful analysis in following stages.

3.2.1 Data Acquisition and Meta Data Collection

The initial phase of the research involved data collection using the Twitter API. This process allowed us to obtain a dataset containing tweet messages relevant to the search terms, along with crucial metadata. This metadata included timestamps, tweet sources, and indicators of truncation or responses. Timestamps were used to analyze temporal aspects, revealing patterns and trends over time. Furthermore, the inclusion of tweet sources and reply indicators provided valuable insights into user engagement dynamics and interaction patterns.

3.2.2 Geocoding and Spatial Contextualization

In order to perform geospatial analysis, it was essential to establish the geographical context of the tweets. The initial dataset contained latitude and longitude data, which was subsequently enhanced through the utilisation of the Nominatim geocoder from the 'geopy.geocoders' module. The geocoder utilised coordinates to assign location information, such as addresses or place names, to each tweet. However, further cleaning was required in order to confirm the accuracy of the location information. In order to accomplish this, reverse geocoding operations were conducted utilising the OpenCage

Geocoding API. The implementation of this procedure significantly enhanced the spatial context of the information, facilitating a full comprehension of the distribution of tweets across various locations within the UK.

3.2.3 Data cleaning and Pre-processing

In order to provide comprehensive examination of the dataset, it was imperative to perform data cleaning as a crucial preliminary procedure. To ensure the integrity of the data, we removed informal phrases, emoticons, symbols, hashtags, and URLs from the tweets. These factors frequently introduce both noise and ambiguity, which might possibly impede the accuracy of later analysis (Otero et al., 2021). Furthermore, the process of data cleaning involved the elimination of grammatical contractions, abbreviations, and acronyms that are frequently employed on social networking sites. The implementation of this meticulous cleaning methodology enhanced the dataset's appropriateness for thorough analysis, guaranteeing uniformity in the utilisation of language throughout the tweets.

3.3 Text Mining Processes

The objective of the study is to obtain a thorough understanding of discussions pertaining to water contamination on the Twitter platform within the context of the UK. In order to accomplish this objective, sophisticated methodologies including word frequencies, topic modelling (LDA and LSA) and sentiment analysis were adopted to analyse the twitter data. These methodologies facilitate the retrieval and examination of many topics being discussed, together with the emotional sentiment linked to tweets.

The rationale for selecting LDA and LSA as the chosen methodologies for the present study is based on many key factors found on the literature. LDA is a widely recognised and extensively studied topic modelling technique that has been among the various alternatives for topic models that have been examined, LDA emerges as the favoured selection for a number of compelling reasons (Lin et al., 2016). LDA operates within a probabilistic framework, which enables the incorporation of uncertainty and the representation of the underlying uncertainty inherent in the data. As confirmed by Kalepalli et al. (2020) inherent probabilistic characteristics of LDA contribute to its enhanced resilience in effectively managing textual material that is characterised by noise and diversity. Furthermore, LDA offers superior topic interpretability in comparison to alternative models such as Singular Value Decomposition (SVD). (Kalepalli et al., 2020; Mohammed and Al-augby, 2020; Sorour et al., 2014).

LSA functions by generating a term-document matrix, whereby each row corresponds to a term, each column represents a document, and the values indicate the frequency of term occurrence. By utilising singular value decomposition (SVD), LSA effectively decreases the dimensions of the given matrix, hence revealing latent semantic patterns and relationships that are present within it (Sorour et al., 2014). LSA is able to successfully identify synonymous terms and uncover the contextual proximity of words by converting the high-dimensional space into a lower-dimensional one. This capability allows LSA to comprehend the semantic meaning of text beyond just superficial co-occurrence. This methodology is utilised in various domains such as information retrieval, text summarisation, and document clustering, among other areas. In this study LSA was utilised, alongside LDA to enhance the comparison between these two methods (Wang and Zhu, 2017).

3.3.1 Wordclouds

A word cloud, alternatively referred to as a tag cloud or text cloud, is a visual depiction of words derived from a collection of texts, wherein the size of each word is proportional to its frequency or

significance within the text. Word clouds offer a rapid and easily comprehensible summary of the most frequently used phrases, enabling analysts to discern crucial subjects, patterns, and attitudes within a given collection of data. Within the domain of social media data mining, word clouds have experienced a surge in popularity owing to their capacity to succinctly represent extensive textual data in a visually comprehensible manner (Kabir et al., 2018). Word clouds were incorporated in this study to visualise the frequency and prominence of words and analyse the topic models used in this research.

3.3.2 Sentiment Analysis

Sentiment analysis, alternatively referred to as opinion mining, entails the systematic procedure of ascertaining and classifying the emotional expressions conveyed within a given textual composition (Feldman, 2013). This study utilised sentiment analysis to evaluate the polarity of feelings expressed in tweets pertaining to water pollution, categorising them as positive, negative, or neutral. This methodology facilitates the assessment of the general public's sentiment towards water pollution matters, the identification of probable origins of concern or discontent, and the comprehension of the public's perception of pertinent policies and initiatives (Otero et al., 2021).

Sentiment analysis algorithms in this study employed approaches from the field of natural language processing (NLP) to examine textual data and discern words and phrases that convey sentiment. The determination of sentiment in a tweet is afterwards based on the frequency and contextual usage of the words present. Positive feeling is indicative of a favourable disposition, whilst negative sentiment signifies discontent or disapproval. Conversely, neutral sentiment denotes a lack of pronounced emotional expression (Hasan et al., 2019).

Through the utilisation of sentiment analysis on tweets, researchers are able to gain a deeper understanding of the emotional reactions exhibited by the general public towards occurrences of water pollution. Additionally, this approach allows for a more comprehensive assessment of the efficacy of ongoing endeavours aimed at addressing pollution concerns. Furthermore, it enables the identification of instances where popular sentiment may deviate from official positions on the matter.

3.3.3 Topic Modelling - Implementing LDA and LSA in Python

Topic modelling is a highly effective methodology employed to reveal underlying themes or subjects within a collection of texts, such as tweets. Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are frequently employed approaches in the field of topic modelling (Tong and Zhang, 2016). LDA is a probabilistic model that posits the notion that every document, such as a tweet, is composed of a combination of multiple subjects. Furthermore, each topic is represented as a probability distribution across words. The objective of LSA is to discern latent topics and their respective word distributions from a collection of tweets (Onan et al., 2016; Otero et al., 2021).

In the present work, LDA was utilised to extract prominent topics or issues found in tweets pertaining to water pollution. Additionally, LSA was utilised to ensure a comprehensive comparison between the two methodologies.. Through the process of identification and categorization, researchers have been able to acquire a more profound comprehension of the predominant subjects pertaining to water pollution in the UK as addressed on the Twitter platform. For example, LDA and LSA has the potential to uncover various themes related to water pollution, including but not limited to "marine pollution," "plastic waste," and "river pollution." This analysis can offer significant insights into the concerns and priorities of the general population concerning water pollution.

3.4 Data Visualisation

The data obtained from topic modelling and sentiment analysis, underwent processing and were afterwards represented through the utilisation of graph networks and heat maps. The utilisation of data visualisation serves as a potent instrument for academics and policymakers, enabling them to comprehend intricate patterns and trends in a more intuitive manner. Graph networks provide a visual depiction of the interconnections of various subjects, emotions, and picture categorizations, facilitating the identification of potential associations and valuable observations.

In contrast, heat maps serve to visually represent spatial patterns and fluctuations, effectively illustrating the dispersion of tweet subjects, emotional tones, and visual attributes among various geographic areas of the United Kingdom. The aforementioned visualisations offer a comprehensive perspective on the prevailing public sentiment on water pollution, the specific regions where worries are most concentrated, and the efficacy of implemented pollution control measures.

Hence, several techniques such as topic modelling, sentiment analysis, LDA, and LSA has contributed to a more comprehensive comprehension of debates pertaining to water pollution on the social media platform Twitter, specifically within the UK. The utilisation of these sophisticated methodologies facilitated the retrieval of underlying themes, classification of emotional tone, and discernment from data that is not in textual form. Through the utilisation of data visualisation techniques, researchers were able to get significant and informative perspectives on public sentiment, the ramifications of water contamination, and the potential policy consequences. The use of this comprehensive approach facilitates a deeper understanding of the patterns and trends of water pollution, so offering a valuable contribution to the process of informed decision-making and effective environmental management within the nation.

4 Results and Discussion

4.1 Spatial Distribution of Tweets

The outcomes of conducting a reverse geocoding operation on the original dataset, resulting in the precise determination of the tweet locations is illustrated in figure 6. Through this visualisation, the spatial distribution of tweets across the UK is portrayed, offering valuable insights into the geographic spread of user-generated content. By employing reverse geocoding techniques, the raw tweet data, initially consisting of coordinates or vague location information, has been transformed into specific locations within the UK. This process enhances the understanding of where the tweets originated from, enabling researchers to discern patterns, trends, or regional variations in the data set.

Twitter Data on UK Region Map (After Correction)

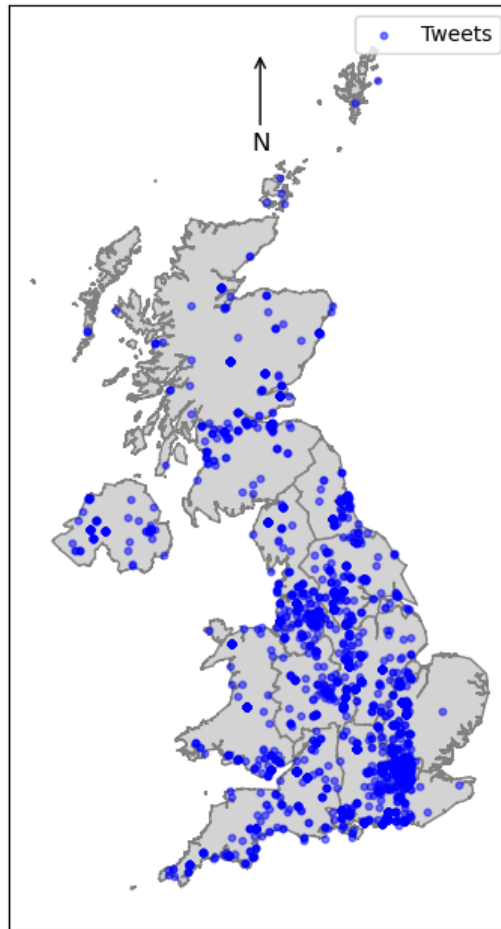


Figure 6. Spatial Distribution of Water Pollution Tweets in the UK, between 18/06/2023 and 26/06/2023

The hourly temporal pattern of tweets is illustrated in figure 7, shedding light on the frequency of tweet activity throughout the day. Interestingly, the data reveals a substantial increase in the number of tweets during off-peak hours, suggesting a distinct pattern and a prime period for tweeting. However, it is important to note that while this pattern is informative in understanding the temporal behavior of tweets, it does not directly explain the underlying nature of bias, the specific categories of tweets, or the topics being discussed.

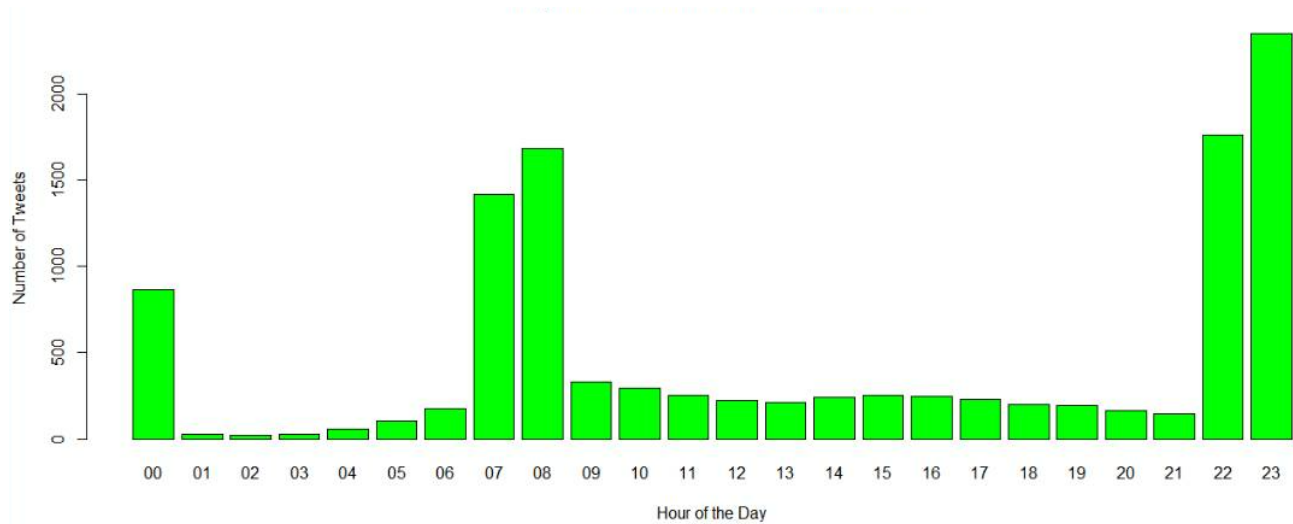


Figure 7. Temporal Patterns of Water Pollution Tweets

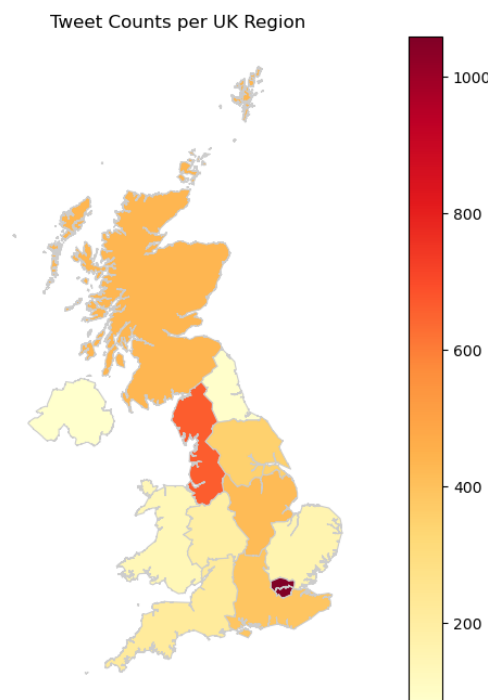


Figure 8. Heatmap of Tweets in the UK

The distribution of tweet counts across different regions within the UK is depicted in figure 8. Nevertheless, it is important to note that the figure lacks normalisation and fails to account for population density in the present phase of the analysis. In doing a more comprehensive analysis, it is imperative to consider the variable of population density.

4.2 Word clouds and Frequency of Words

The significance of regional word clouds in the context of sentiment analysis about water pollution in the UK cannot be overstated, as they play a crucial role in enabling a more nuanced understanding of public attitudes. Through the analysis of localised word patterns, these clouds reveal attitudes that are distinctive to particular regions, so providing insights into the varied perspectives held across the nation. The level of granularity provided allows policymakers and environmental authorities to customise responses based on specific area considerations. Moreover, this study facilitates the identification of the most prominent words in different regions in the UK, hence enabling the implementation of focused communication strategies to successfully address problems. The impact of water pollution might exhibit variability based on the specific locality. The utilisation of regional word clouds serves to enhance comprehension of the emotions experienced by individuals, hence facilitating a more precise and proactive approach towards addressing environmental concerns. Moreover, this technique fosters increased public participation and trust.

The regional word clouds and frequencies of words appeared in tweets are depicted in the figures 9-20 for the regions East, East Midlands, West Midlands, North East, North West, South East, South West, London, Yorkshire and Humber, Scotland, Wales and Northern Ireland respectively.

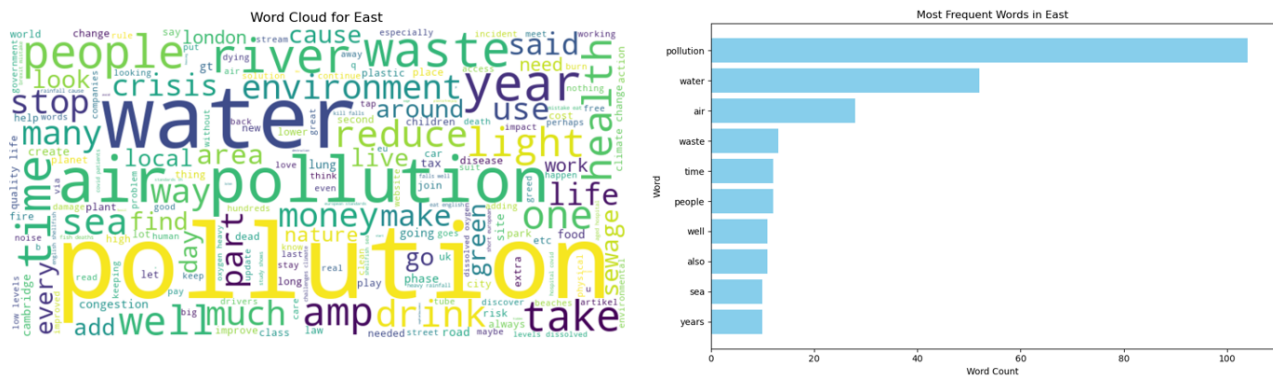
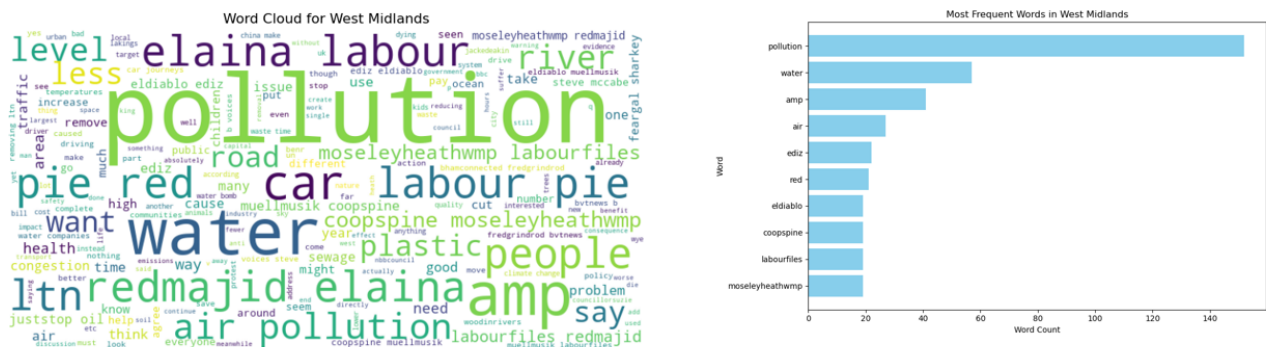
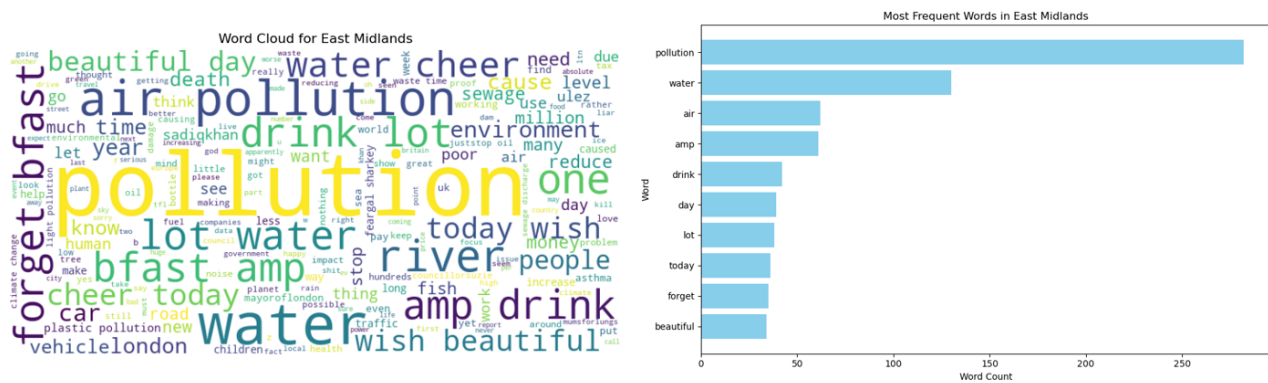
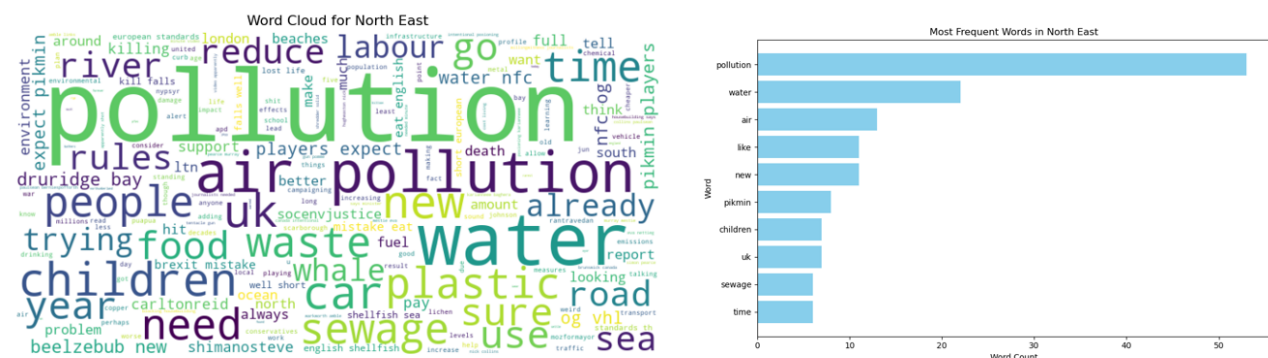


Figure 9. Word Cloud & Word Frequencies of Water Pollution Tweets in UK East Region

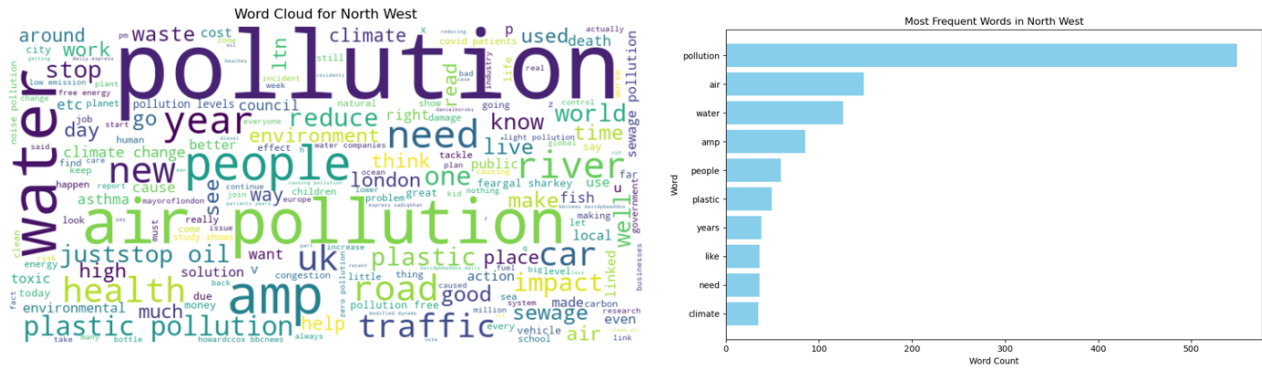
Within the geographical setting of the Eastern area in the United Kingdom, a comprehensive examination of notable terms utilised in tweets revealed a noticeable emphasis on matters pertaining to pollution and issues relating to water. In addition to the core concepts of "pollution" and "water," the inclusion of keywords such as "air," "waste," "time," and "people" offers significant analytical perspectives on the underlying characteristics of water pollution concerns. The word "air" implies a possible correlation between air quality and water contamination, as airborne pollutants have the ability to deposit into aquatic environments. The term "waste" may signify apprehensions regarding the adverse impact of inadequate disposal methods on water systems. The term "time" may allude to the temporal aspects of pollution concerns, suggesting the existence of longstanding difficulties that necessitate care. Lastly, the incorporation of "individuals" indicates the socio-environmental ramifications of water pollution, alluding to issues related to public health and communal welfare. The analysis indicates that there is a complex challenge including multiple aspects of contamination. This suggests that there is a requirement for comprehensive and sustainable measures to effectively tackle water quality problems in the Eastern region.



In the East Midlands and West Midlands regions, the prevalent terms, aside from "pollution" and "water," included the word "air." This implies the possibility of air pollution perhaps as a resultant effect of the underlying water pollution, particularly in the East Midlands region. Nevertheless, the absence of major terms such as 'waste' suggests a favourable perception regarding the management of garbage in the respective areas.

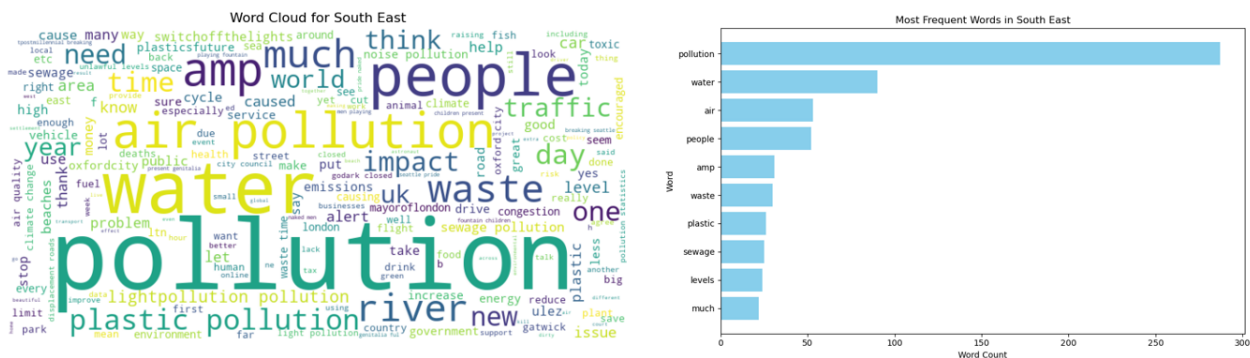


Upon analysing the word clouds pertaining to the North East region, it becomes evident that the term 'sewage' has exhibited a significant frequency. This observation suggests that sewage has emerged as a prominent issue within the region, hence directly impacting the water pollution levels in the area.



The word cloud pertaining to the North West region prominently highlights a notable concentration of the term "plastics," suggesting a substantial occurrence of talks relating to plastic substances within the region. The frequent usage of the term "plastics" suggests a widespread adoption and potential apprehension surrounding the utilisation of plastic-based products within the area.

The conspicuous prevalence of the term "plastics" in the word cloud of the North West region indicates a pressing necessity for collaborative endeavours aimed at tackling plastic waste management and fostering sustainable practises within the locality.



The word cloud for the South East region reveals a notable presence of phrases such as "waste," "plastics," and "sewage," which suggests that waste management and environmental pollution are prominent problems in this area. These phrases collectively indicate a geographic area that is facing challenges pertaining to the management of garbage disposal, plastic pollution, and sewage. The prevalence of "waste" signifies an emphasis on the comprehensive administration of abandoned substances, potentially suggesting difficulties in the appropriate handling and re-utilisation methods. The concept of "plastics" highlights the increasing recognition of the adverse impacts of plastic pollution on aquatic environments and ecological systems, indicating a necessity for enhanced practises of sustainable consumption and waste mitigation. In the meantime, the incorporation of the term "sewage" serves to underscore apprehensions regarding the treatment of wastewater and its potential ramifications on the quality of water resources. The word cloud composition indicates the need for comprehensive measures that encompass trash generation, plastic pollution, and sewage treatment in order to minimise the water pollution in the South East region.

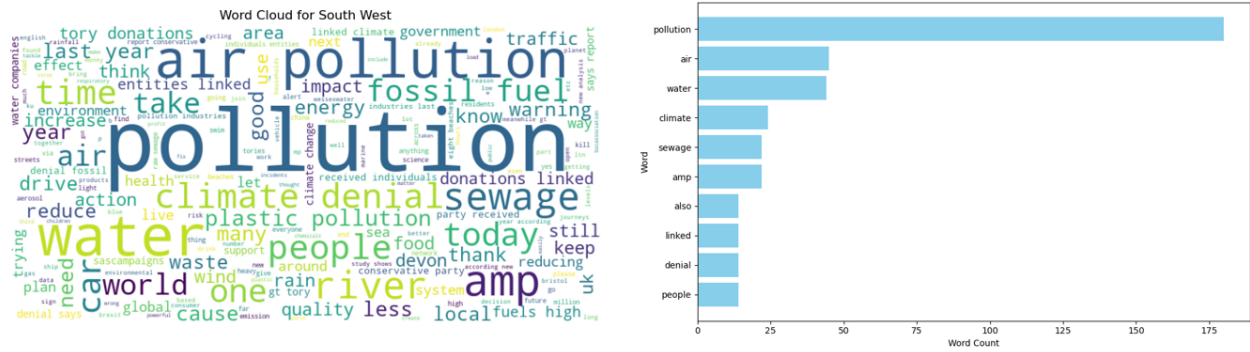


Figure 15. Word Cloud & Word Frequencies of Water Pollution Tweets in UK South West Region

In contrast to the South East region, the South West region just encompasses the term 'sewage', so implying that there is ongoing discourse among individuals over the sewage system within this area. This suggests that, in contrast to the South East region, the general people in this area does not prioritise the issues of plastics and waste.

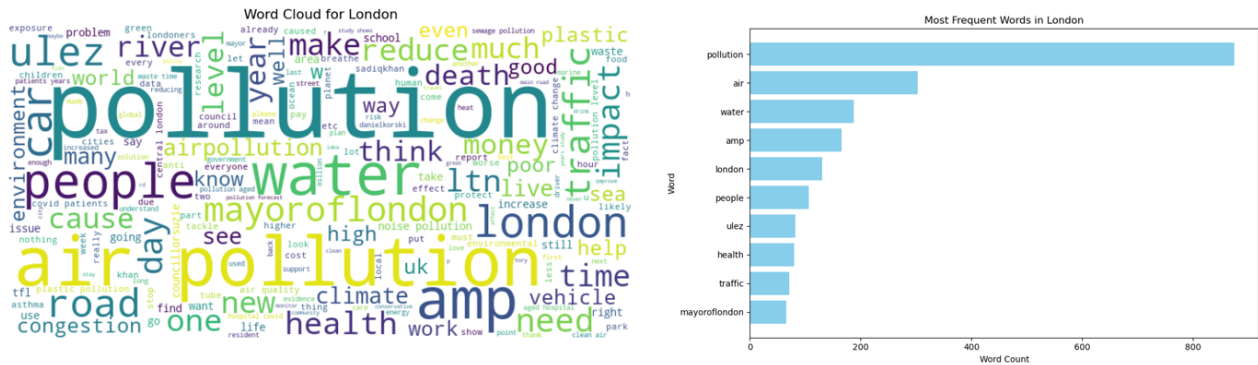
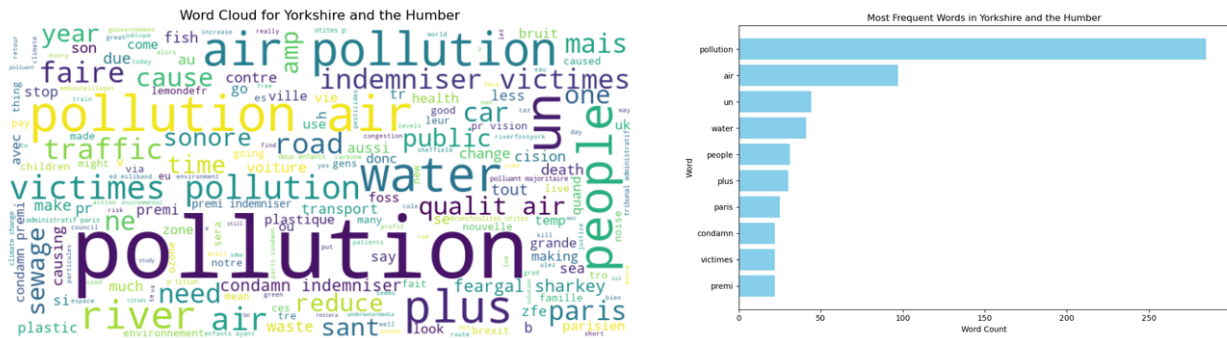


Figure 16. Word Cloud & Word Frequencies of Water Pollution Tweets in London

Upon analysing the discourse on Twitter pertaining to London, it becomes evident that the word cloud prominently features the term "pollution," with a notable focus on "air." This observation suggests a significant level of concern regarding the issue of air quality. Significantly, the term "air" is given priority over "water," indicating a heightened emphasis on air pollution in comparison to water pollution. This observation indicates an urgent requirement to tackle the air quality concerns in the region. Furthermore, the use of terms such as "health" and "traffic" highlights the wider ramifications of air pollution, including its possible effects on public health and the occurrence of traffic congestion. The word cloud exhibits a substantially lower degree of prominence for the term "water," and the absence of associated terms may indicate a more suppressed level of discourse regarding water contamination. This highlights the pressing need to tackle air quality issues in London, so motivating a contemplation of holistic approaches that involve the reduction of both air and water pollution.



Similar to the regions of London and North West, the Yorkshire and Humber region likewise has a higher frequency of the word "air" compared to "water," indicating a greater emphasis on discussions related to air pollution. Furthermore, it is worth noting that terms such as "waste" and "sewage" and "plastics" were conspicuously absent from the discussions, thereby providing additional evidence to support the aforementioned remark.

Upon investigating the twitter conversations surrounding Scotland, the prominent lexical items "pollution," "air," "low," and "quality" within the word cloud collectively indicate a notable focus on apprehensions over the state of air quality. The term "pollution" encompasses a wide range of environmental concerns; nevertheless, the specific mention of "air" and the emphasis on "quality" indicate a particular interest in the monitoring and mitigation of air quality levels.

The absence of the term "water" implies a favourable indication of less water pollution within the area.

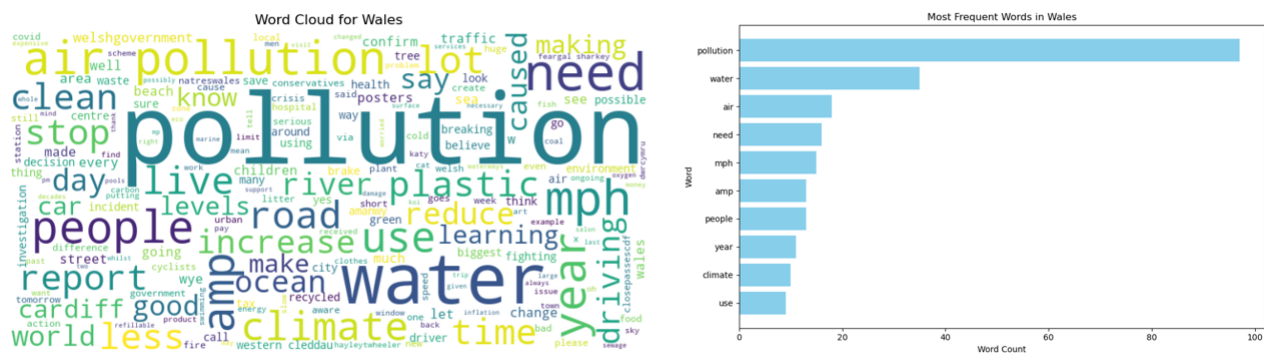


Figure 19. Word Cloud & Word Frequencies of Water Pollution Tweets in Wales

Upon analysing the twitter discussions pertaining to Wales, it becomes evident that the prevailing terms, including "pollution," "water," "climate," and "people," as depicted in the word cloud, collectively construct a narrative that encompasses a comprehensive viewpoint on environmental issues. The term "pollution" implies a broad recognition of the deterioration of the environment, which may include many forms of contamination such as air and water pollution. The salience of "water" underscores a distinct emphasis on matters pertaining to water, signifying apprehensions regarding water quality and maybe mirroring the significance of water resources in the locality. The incorporation of the term "climate" denotes a comprehension of the wider environmental framework, implying an acknowledgment of the interdependence among diverse ecological elements. Lastly, the inclusion of "people" highlights the anthropogenic aspect of these ecological dilemmas, accentuating the potential consequences for nearby societies and the well-being of the general public. The word cloud representation presented here indicates a collective recognition of several environmental aspects in Wales, encompassing issues such as water quality, broader climate considerations, and concerns related to human well-being.

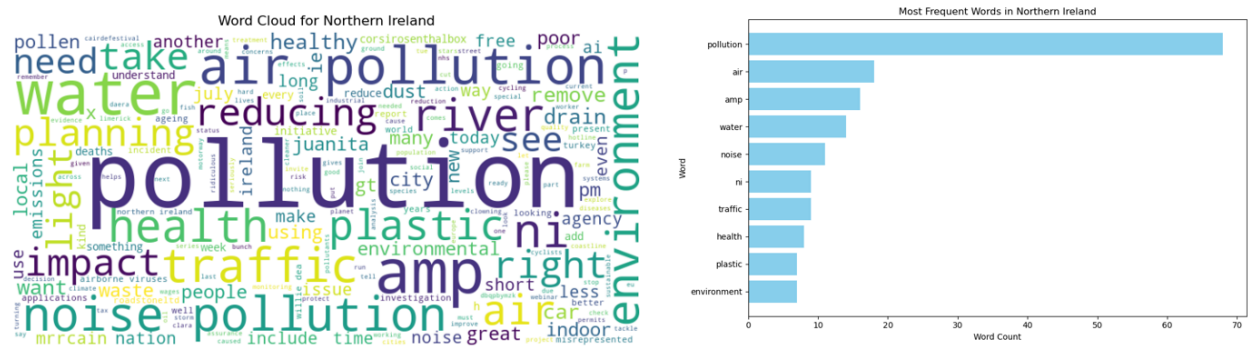


Figure 20. Word Cloud & Word Frequencies of Water Pollution Tweets in Northern Ireland

On examining the conversations pertaining to Northern Ireland, the prominent lexical items "air," "water," "noise," and "plastic" depicted in the word cloud collectively provide a thorough portrayal of the environmental issues covering both air and water pollution. The term "air" denotes a concentrated emphasis on matters pertaining to air quality, indicating an increased consciousness of potential substances that may impact the atmosphere. The incorporation of the term "water" serves as a symbolic representation of a concurrent focus on the quality of water, so implying apprehensions regarding the detrimental effects of pollution on aquatic environments. The term "noise" might evoke concerns

pertaining to noise pollution, hence emphasising the multifaceted character of environmental challenges. Moreover, the inclusion of the term "plastic" raises worries on the issue of plastic pollution, so demonstrating an understanding of its detrimental effects on ecosystems on land as well as in the water. The aforementioned depiction emphasises the significance of tackling various aspects of pollution in Northern Ireland, emphasising the need for comprehensive approaches that encompass the maintenance of air and water quality, mitigation of noise pollution, and effective management of plastic trash.

The examination of Twitter data pertaining to water pollution in the UK yields significant insights when employing word cloud analysis techniques. The analysis of word frequency provides insight into the prevailing themes and issues, hence illuminating the public's concerns (Kabir et al., 2018). The analysis of word sentiment enables the identification of the emotional tone present in dialogues, encompassing emotions such as alarm, rage, or hopefulness. The identification of notable locations regions or pollutants referenced in the text aids in the precise determination of certain areas of focus (Kaptein, 2012). A clear example for this could be seen through the results of this study where some regions have more concern on plastics where as few other regions have discussed more about sewage.

The act of monitoring popular phrases over a period of time allows for the observation of evolving narratives and changing priorities. Through the comparison of word clouds across several time periods, discernible visual patterns emerge, effectively illuminating the dynamic nature of evolving trends (Kabir et al., 2018). These studies provide policymakers, companies, and environmental organisations with the necessary tools to make well-informed decisions, tackle urgent challenges, and customise their communication strategies. Nevertheless, it is imperative to adopt a nuanced perspective when considering the findings, as there may be inherent demographic biases present among Twitter users and the subtle complexities of language usage must be taken into account (Valle-Cruz et al., 2017).

4.3 Sentiment Analysis

This section provides an overview of the predominant sentiments expressed in tweets across the entire country and a comparative analysis specific to each region of the UK. The distribution of positive, negative and neutral twitter data sentiment about water pollution as mined from Twitter data are presented through figures 21 and 22.

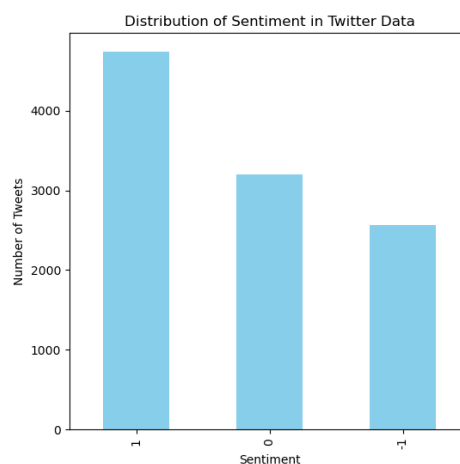


Figure 21. Distribution of Twitter Data Sentiment

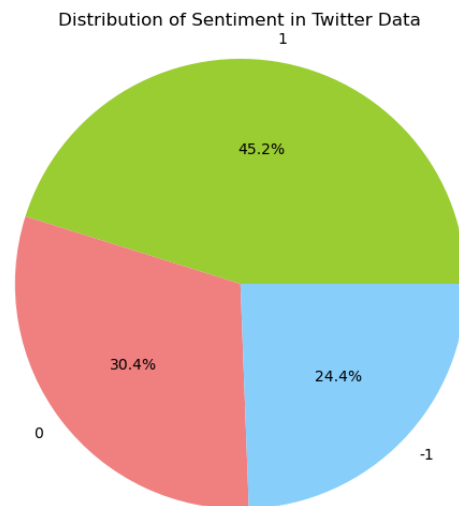


Figure 22. Distribution of Twitter Data Sentiment:Pie Chart

The analysis of sentiment within Twitter data concerning water pollution in the UK has revealed a wide array of thoughts and emotions being conveyed by users. This diversity of sentiments is evident in the data, as illustrated in Figures 21 and 22. Notably, the analysis indicates that a significant portion of tweets, approximately 45.2%, exhibit a positive sentiment (1). This finding underscores the favorable stance taken by a considerable number of users towards the various initiatives and efforts aimed at addressing water pollution in the UK. These figures provide compelling evidence that a substantial proportion of individuals hold a positive perception of the strategies and actions implemented to combat water pollution. The noteworthy aspect here is the substantial agreement across a wide spectrum of users, highlighting a general consensus regarding the effectiveness and importance of the measures put in place to mitigate pollution and ensure the preservation of the quality and sustainability of water resources. This collective recognition of the undertaken measures underscores a widespread acknowledgment of the ongoing endeavors to tackle the issue of water pollution (Otero et al., 2021).

In contrast, there is a noticeable divergence in the statistics, with approximately 24.4% of the analyzed tweets displaying a discernibly negative sentiment marked as (-1). This negative sentiment is indicative of prevailing apprehensions and discontent related to the pressing issue of water pollution. Within this segment of Twitter users, there appears to be a concerted effort to shed light on instances of insufficient pollution control, resulting ecological harm, and an apparent lack of effective regulatory protocols. The viewpoints and opinions conveyed by the individuals underscore the urgent requirement for substantial improvements in environmental policies and the implementation of more stringent enforcement measures (Ahuja and Dubey, 2017). These steps are deemed essential in safeguarding and preserving our invaluable water resources.

It's important to note that a substantial portion of the tweets, precisely 30.4%, exhibit a neutral sentiment falling under the category of 0. This observation implies that these tweets capture a sense of impartiality and lack of bias in relation to the discussed topic. The content within these tweets might encompass objective information, ongoing events, or balanced discussions that steer clear of any overt emotional influences. In sentiment analysis, a neutral sentiment refers to a state in which the language or material does not exhibit any identifiable emotional tone or inclination towards either positive or

negative sentiments (Sahayak et al., 2015). The language employed in the discourse is indicative of objectivity, devoid of any overt expression of strong emotions or personal viewpoints pertaining to the topic water pollution within the region.

4.3.1 Regional Sentiments by Location

The regional sentiment scores for the regions in the UK were calculated in this study and the graphical representation of these regional sentiments are illustrated in the figure 23.

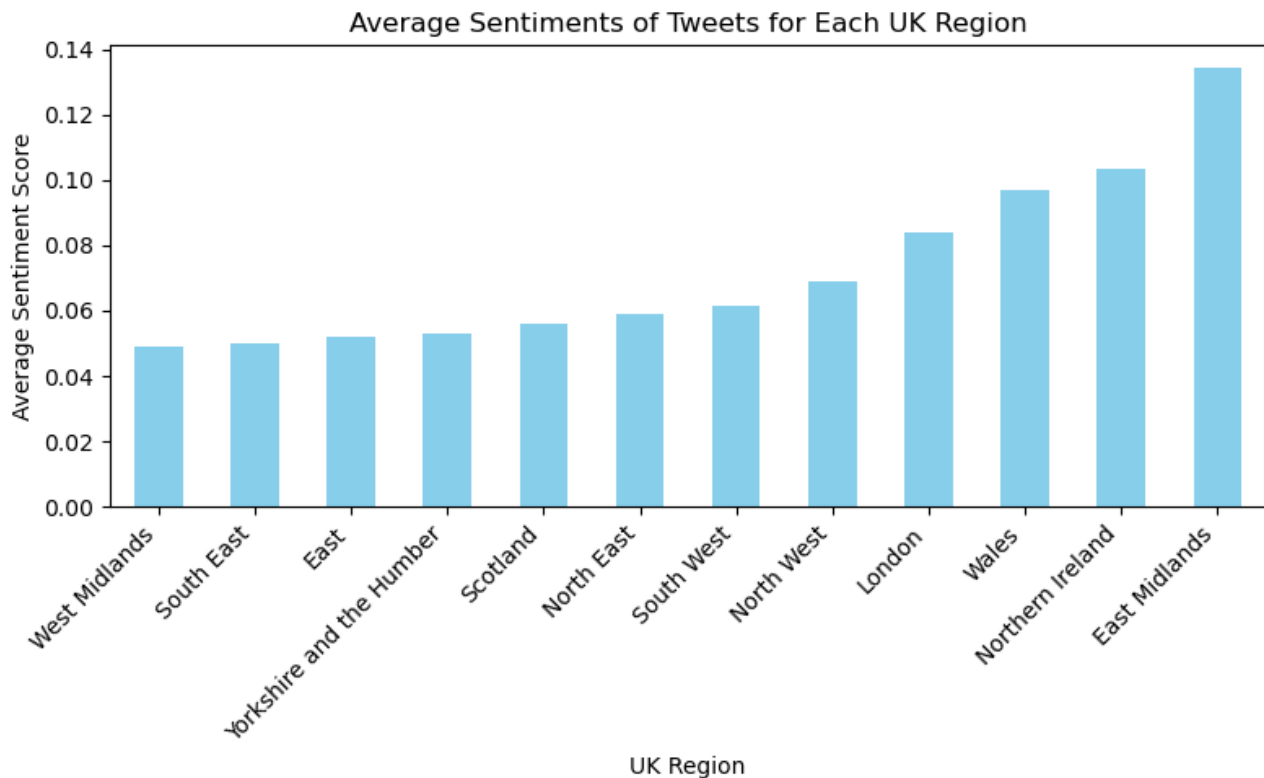


Figure 23. Average Regional Sentiment scores in the UK

In the context of sentiment analysis, the sentiment values exhibited along the y-axis hold significance as they represent the average sentiment scores derived from tweets associated with different regions within the United Kingdom. These sentiment scores are typically distributed across a numerical scale that ranges from -1 to 1. In this scale, negative values are indicative of a pessimistic or adverse sentiment, while positive values symbolise optimistic or positive emotions. Additionally, scores close to the midpoint of 0 indicate a neutral sentiment stance. Conversely, the x-axis of the graph serves as a representation of the diverse geographical segments that comprise the UK. Each individual bar portrayed on the graph is a visual representation of a specific geographical location. Notably, the vertical dimension of the bar graph conveys a significant piece of information. It reflects the average emotional score assigned to tweets originating from the corresponding geographic region. This score is a quantitative measure of the prevailing emotional tone within that particular area, thereby shedding light on the overall sentiment landscape across the UK's different regions.

In the graph, each bar's vertical dimension symbolises the average emotion score extracted from tweets related to a specific geographical area. Taller bars on the graph correspond to a heightened level of positive sentiment, illustrating a stronger positive outlook. Conversely, shorter bars indicate a more pronounced negative attitude, reflecting a greater degree of negativity. If a bar is situated near

the zero mark on the graph, it indicates that the tweets originating from that particular region tend to showcase a relatively balanced and neutral emotional tone.

The sentiment analysis results reveal a coherent and persistent trend that mirrors the insights gleaned from the distinct word clouds representing diverse geographical areas. The discernible fluctuations in sentiment scores across these regions serve as telltale signs of the existence of diverse attitudes and emotional responses in relation to the water pollution issue. Specifically, the localities encompassing the West Midlands, South East, East, and Yorkshire and Humber manifest relatively diminished sentiment scores, implying a proclivity towards maintaining a more equilibrated and impartial demeanor regarding the concern of water pollution. In these particular areas, discussions pertaining to this subject might transpire with a moderated degree of apprehension, characterised by a conscious effort to avoid overly impassioned reactions.

On the other hand, it becomes apparent that in places like Wales, Northern Ireland, and East Midlands, conversations revolving around water pollution tend to evoke distinctly positive reactions. These regions exhibit sentiment scores that eclipse the 0.10 mark, signifying a prevailing tendency towards favorable emotional engagement. This numerical threshold takes on added significance as it signifies the presence of prominently agreeable sentiments or a highly hopeful vantage point. This positivity-infused sentiment score provides valuable insight into the collective emotional disposition of these areas when addressing the issue of water pollution, suggesting an encouraging and optimistic perspective in these conversations.

In the context of positive emotions, it becomes apparent that scores surpassing the threshold of 0.10 carry substantial implications, pointing towards noteworthy progress, ambitious undertakings, or notable achievements in the ongoing battle against water pollution within the specified regions. This deduction underscores the existence of a discernible relationship between the sentiments expressed by the general public and the tangible execution of affirmative actions targeted at mitigating the issue (Otero et al., 2021). In essence, the statistical correlation between the sentiment levels and the implementation of effective measures signifies that public opinion plays a role in motivating and shaping the constructive initiatives that are being deployed to combat water pollution in these areas.

Conversely, sentiment ratings that exceed the 0.10 threshold also indicate a heightened degree of positive reactions. These elevated sentiment levels serve as clear indicators of strong support, validation, and acknowledgment of the concerted efforts directed towards combatting water pollution (Karami et al., 2020). The sentiment scores exceeding this benchmark signify a profound and widespread endorsement of the initiatives that have been set in motion to address the issue. This suggests that the strategies, actions, and policies aimed at tackling water pollution have garnered not only recognition but also a substantial degree of approval and commendation from the public, underscoring the resonance of these efforts within the community.

The sentiment scores for regions including Scotland, North East, Southwest, North West, and London, fall within the range of 0.06 to 0.08. These scores suggest the presence of a moderate level of sentiment expression within these regions. Specifically, when discussing the topic of water contamination, these geographical areas display a moderate yet discernible degree of positive sentiment in their discourse. However, it is of paramount importance to acknowledge that sentiment analysis provides a sweeping overview of emotional content. To achieve a deeper understanding, it is imperative to engage in qualitative research focusing on individual discussions within distinct geographic regions. By delving into this approach, a more nuanced comprehension emerges, shedding light on the intricate details that underlie emotional responses concerning water pollution (Saldana-Perez et al., 2019).

This study plays a pivotal role in enhancing our grasp of the fundamental causes and contextual

elements that give rise to the emotional reactions documented in connection with water pollution. Through a meticulous examination of specific conversations occurring across various geographical areas, we can unravel the intricate tapestry of factors influencing these emotional responses (Otero et al., 2021). This approach transcends the limitations of sentiment analysis, offering a comprehensive and multifaceted insight into the complex interplay of emotions and environmental concerns.

4.4 Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA)

This section examines the intricate details of public discourse surrounding water pollution in the United Kingdom by employing two distinct topic modelling techniques: LDA and LSA. The aim of this analysis is to extract valuable information from the predominant words linked to the created themes, so enabling a comprehensive comprehension of the discourse surrounding water pollution on the Twitter platform.

4.4.1 Analysis of Latent Dirichlet Allocation (LDA)

The LDA model, which is a probabilistic generative model, has exceptional proficiency in capturing the multifaceted nature of discussion pertaining to water pollution. The results obtained through the LDA analysis of this study are depicted in figures 24 and 25.

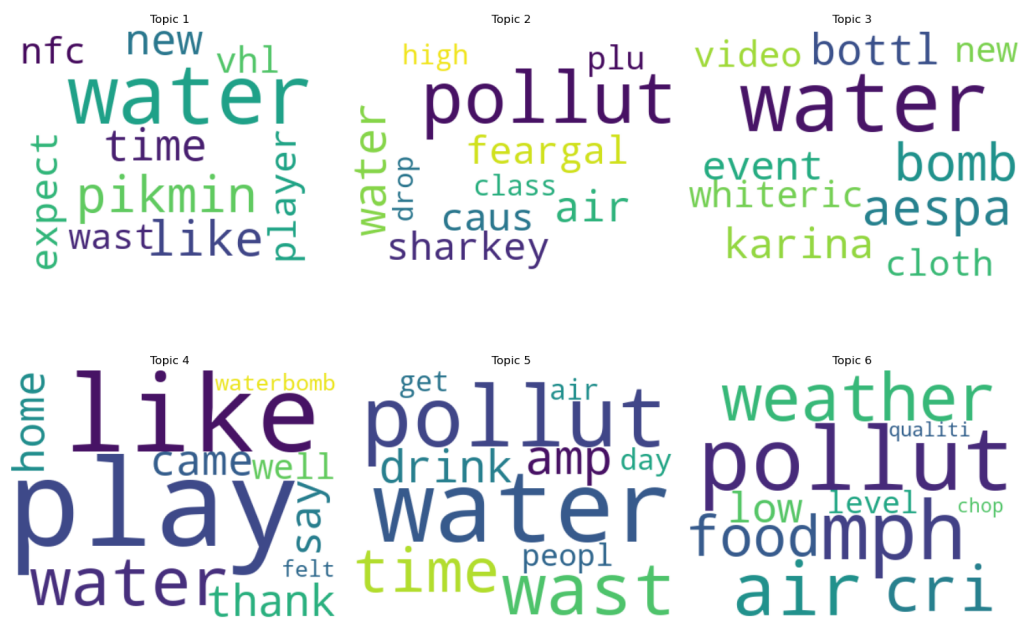


Figure 24. Word clouds for LDA model topics

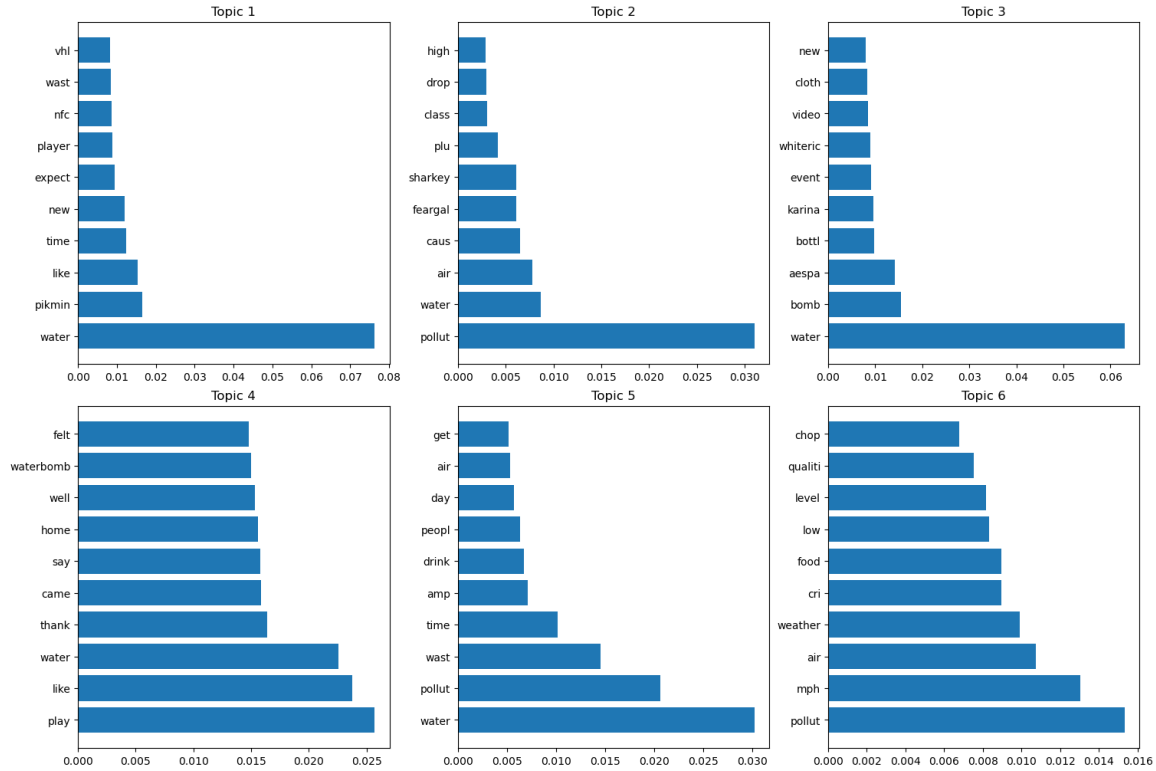


Figure 25. LDA Bar plot showing weighting of words in each topic.

In this context, it can be observed that each bar within the visual representation (figure 24) symbolises a certain word, with the length of each bar serving as an indicator of its relative significance within the relevant topic. The horizontal axis denotes the significance score assigned to each word, while the vertical axis displays the top terms associated with each topic. The aforementioned plots offer a straightforward visualisation of the most impactful words within each topic.

Through the process of perception, this model is able to identify and categorise six unique subjects derived from the Twitter data. The topics discussed in the text are not expressly labelled; however, one can infer them based on the major words used. This effectively reveals the complex and multifaceted nature of public perceptions. The terms - "water", "pollution", "waste", "quality", "drink", "air" and "people" are the most prominent words that were extracted across all topics from LDA model.

The term "water" serves as the central element, representing the primary focus. The topics of "pollution" and "quality" serve as the focal points of the conversations, highlighting the collective apprehension regarding the contaminated condition of water bodies. The incorporation of the term "waste" implies a recognition of the wider ecological consequences associated with water pollution. Moreover, the inclusion of the term "air" implies that the topic has recognised the potential interrelationship between cross-pollution phenomena, wherein water pollution could potentially impact the quality of the surrounding air. The utilisation of the term "people" reveals a crucial aspect of these conversations, namely, the impact on human health and overall welfare. The correlation between the consumption of drinks and pollution highlights the apprehensions around the safety of drinking water, suggesting that individuals are concerned about the availability of clean and unpolluted water resources. The amalgamation of these concepts exemplifies a holistic perspective on the manner in which water pollution impacts individuals' lives, intricately linking environmental, health, and societal aspects.

The presented visual depiction in figure 26, as a heatmap, illustrates the distribution of textual mater-

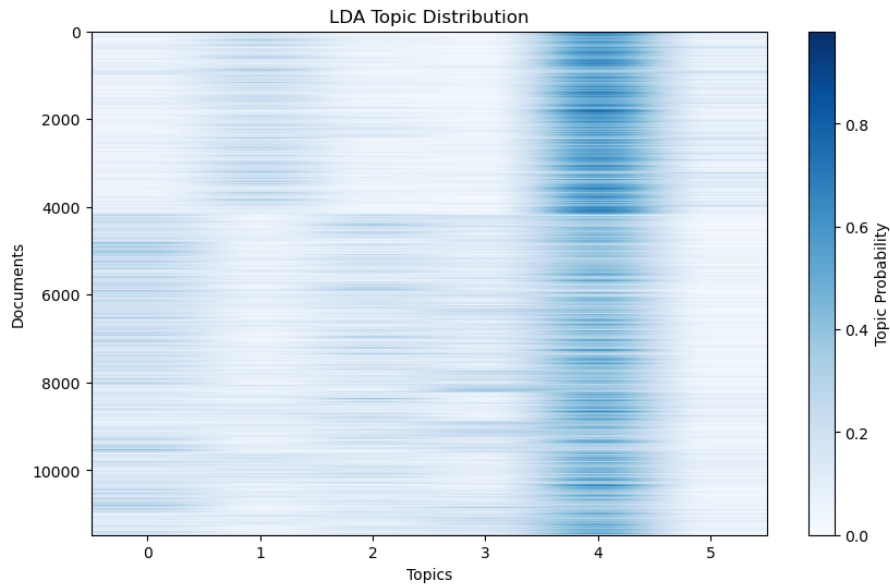


Figure 26. Topic Distribution of LDA model

ials or documents across various themes related to water pollution in the UK. Within this illustrated framework, each row represents a distinct document, while the columns are representative of designated themes. The heatmap utilises chromatic intensity in each cell to indicate the level of correlation between a specific document and a particular topic. A cell exhibiting a deeper shade of colour indicates a higher probability of correlation between the document and the subject matter of the respective issue. The utilisation of this visual instrument enables a more nuanced understanding of the current prevalence of various subjects related to water pollution in the UK. This understanding is derived from a collection of documents that have undergone text mining techniques using data obtained from the Twitter platform.

4.4.2 Latent Semantic Analysis (LSA)

LSA is a computational method used to analyse the semantic similarity of words and documents. In this context, the focus is on examining the limitations associated with a narrow representation of information. The word clouds generated by LSA and the frequency of the words been prominent in the model are illustrated in figures 27 and 28.



Figure 27. Word clouds for LSA model topics

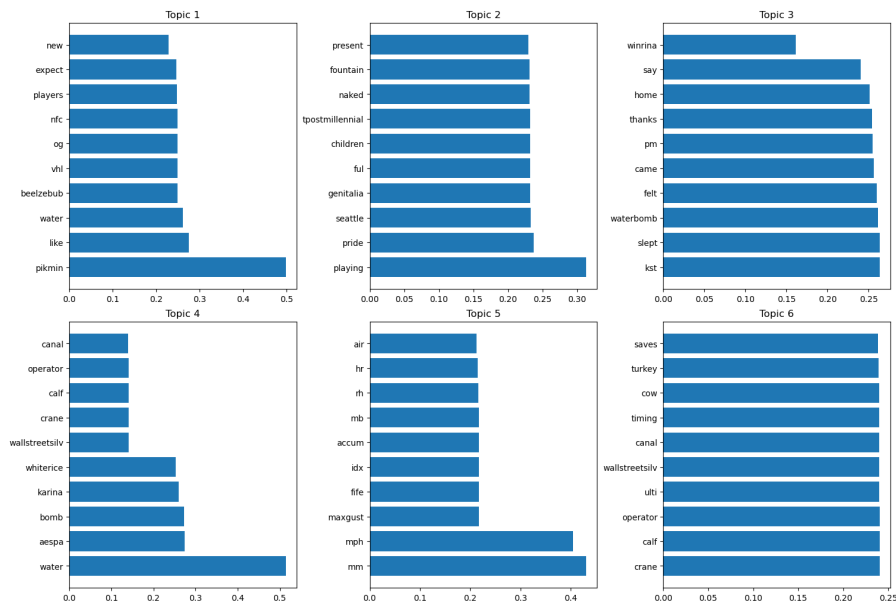


Figure 28. LSA Bar plot showing weighting of words in each topic.

Hence, the LSA analysis of the study provides a narrower viewpoint. When constructing a list of six topics, the scope of related terms is restricted to "water", "people" and "children". The produced topics appears to be ambiguous and lacks a strong connection to water pollution, as observed in the LDA analysis. The choice of "water" as the subject of investigation is unsurprising considering the study's primary emphasis. The terms "people" and "children" elucidate the attention given to populations that are at risk, specifically those who are more prone to experiencing the detrimental consequences of water pollution. Nevertheless, the significant disparity between the profoundness of themes generated by LDA and the restricted extent of findings obtained from LSA prompts inquiries regarding its capacity to encompass the intricacy of public conversation.

Similar to the LDA a topic distribution heatmap was generated through LSA. In the context of LSA, the standard method for representing topic distribution involves the application of a stacked bar chart. In the above visual depiction, each separate bar represents a unique document, with the segments within the bars representing the proportions assigned to certain topics. The vertical length of each segment functions as a visual representation of the significance of a particular issue within the boundaries of the corresponding document. This specific visualisation style enables a thorough understanding of

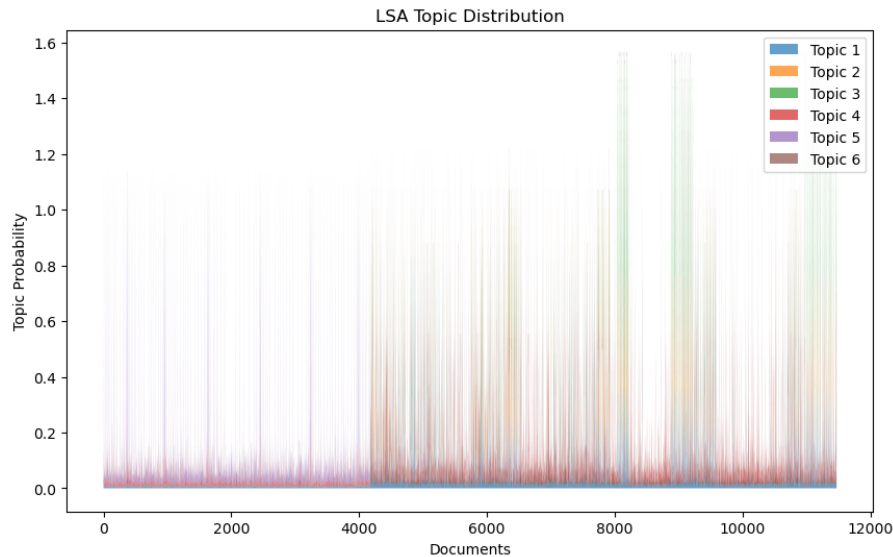


Figure 29. Topic Distribution of LSA model

the structural composition of subjects within individual papers.

The different results seen between LDA and LSA highlight the superiority of LDA within the specific context being considered. The strength of Latent Dirichlet Allocation (LDA) rests in its capacity to uncover a broader range of pertinent phrases, hence facilitating a more comprehensive comprehension of public perceptions. LDA provides a comprehensive representation of the discourse by acknowledging a wide range of areas or topic that embrace not only the central matter of water pollution but also its larger implications such as addressing low water quality and interrelationships between water pollution and air quality. On the other hand, the limiting of LSA to the concepts of 'water,' 'people,' and 'children' can be seen as indicative of a reductionist methodology. The absence of diversity in these topics fails to acknowledge the complex array of issues surrounding water pollution. The lack of phrases such as "pollution," "quality," "waste," and "air" in the topics created by LSA indicates its limited capacity to encompass the diverse aspects of the matter.

The output of LDA comprises a collection of themes, with each subject being represented as a probability distribution over words (Mohammed and Al-augby, 2020). This feature enhances user comprehension and facilitates the categorization of issues by prioritising the utilisation of the most pertinent phrases. LDA algorithm exhibits a high degree of flexibility, allowing for seamless adaptation to diverse text data and corpora. The software possesses the capability to effectively process and analyse extensive data sets, rendering it well-suited for practical scenarios that include huge volumes of data (Kalepalli et al., 2020; Tong and Zhang, 2016).

5 Limitations and Future Research

5.1 Limitations and Challenges

During the course of this research, various challenges have arisen in accessing Twitter data. These obstacles stem from the recent change in ownership of Twitter and corresponding modifications to the Twitter API subscription model. Notably, the previous academic research subscription has been deprecated. Presently, only paid API subscriptions are available, with the introduction of a new academic research subscription pending. Regrettably, this lack of accessibility to the necessary data has impeded the progress of this research. The initial intention was to analyse a dataset spanning approximately 3 years for historical insights. However, the existing paid subscription options, which are both accessible and affordable, only grant access to a 7-days window of data.

Additionally, the research has been hindered by administrative delays within the finance department in acquiring the appropriate dataset subscription. Coupled with this are the financial constraints posed by the unfeasibility of securing the necessary data for the envisaged analysis. These combined challenges have resulted in a need to narrow down the scope of the research. As a consequence, more intricate analyses such as image analysis and secondary data validation, which were integral to the original research plan, had to be curtailed.

Beyond the issues related to data acquisition, further obstacles arise from the secondary data sources targeted via Twitter. For example, the London Water Keepers sewage map, a secondary data source via Twitter, only encompasses the geographical area of London. This geographical limitation appears incomplete when compared to the broader scope of the entire UK, which was intended for validation. Furthermore, the availability of similar data from other agencies on Twitter appears limited, indicating a potential deficiency in coverage.

Another non-Twitter secondary dataset that was considered for validation is the Environmental Agency Water Quality data. However, this dataset is confined to England, presenting a coverage limitation analogous to the previously mentioned sewage map. Another predicament arises from the fact that the parameters or determinands employed for sampling in this dataset have not been explicitly specified. Consequently, the results derived from this dataset lack the necessary normalisation to facilitate the ranking of water quality for each individual sampling point.

The culmination of these obstacles has collectively exerted a substantial impact on the overall scope of the research, especially coupled with time limitations. As a result, poses difficulty in achieving some objectives and questions this research seek to answer. The intricate interplay of these challenges has consequently impeded the project's ability to achieve its intended goals and has necessitated significant adjustments to the research scope and methodology. However, it is important to emphasize that despite these setbacks, the potential of this research to assess water pollution remains intact. The project still lays the groundwork for future research endeavors, offering valuable insights into the intricacies of water quality assessment. While the project had to adapt and curtail some of its initial ambitions due to data access and resource limitations, its contribution to the understanding of water pollution and its implications cannot be underestimated. This experience serves as a stepping stone for the development of more robust methodologies and strategies in future research pursuits within this critical domain.

5.2 Future Research

The current research primarily centers on evaluating sentiments and opinions expressed in tweets concerning water pollution in order to gain a better understanding of people's perceptions about the causes of water pollution and their major concerns. However, it's worth noting that these tweets may pertain to water pollution incidents occurring anywhere globally. Therefore, a more advantageous future approach would involve pinpointing the precise geographical location of the events mentioned in these tweets. This approach would enable us to accurately correlate individuals' opinions regarding water pollution concerns and causes with specific incidents. Consequently, this alignment could facilitate the targeted deployment of mitigation measures and efficient resource management strategies, all contributing to the pursuit of a net-zero-positive future.

Futhermore, in the pursuit of attaining the highest level of accuracy and reliability in utilizing the insights extracted from these primary data sources, it is essential to incorporate a comprehensive validation process. This validation endeavor encompasses the integration of secondary data sources, aiming to establish meaningful correlations between pollution incidents and the prevailing land use patterns within the geographical area. Additionally, it seeks to discern any discernible relationships between water pollution events and the operational activities of water companies operating within a specified region. By undertaking this rigorous validation approach, we can offer a more robust assessment of the underlying causes behind a specific water pollution incident. It serves as a crucial step in corroborating the perspectives and concerns of the entire local community affected by the incident. This multifaceted validation process not only enhances the credibility of our findings but also promotes a more holistic understanding of the environmental factors and human activities contributing to water pollution occurrences.

In the long-term trajectory of this research, the overarching vision of this research extends towards a comprehensive approach that involves the amalgamation of diverse online data sources creating a data fusion. This would encompass a wide array of resources, ranging from web news articles to a multitude of social media platforms. The central objective is to devise an automated process that transforms this extensive collection of unstructured data into VGI. This holistic transformation holds the potential to serve as a unified platform for inquiries, offering a valuable repository of spatial and temporal data. Such an innovation could be instrumental in effectively managing water pollution across the UK, aligning with the strategic goal of achieving a net-zero-plus future.

Moreover, the proposed approach carries the promise of far-reaching implications. Beyond its immediate application in water pollution management, it has the capacity to inform diverse research domains. By fostering dynamism, the envisioned platform could evolve into a responsive tool capable of addressing an array of environmental emergencies such as floods, fires, snowstorms, and air pollution. This potential expansion underscores the adaptability and significance of the proposed framework, positioning it as a multifaceted solution for a spectrum of pressing challenges. As the research progresses, its contributions are poised to resonate not only within the domain of water pollution but also as a catalyst for broader advancements in environmental research and crisis mitigation.

6 Conclusions and Recommendations

In the current era characterised by the abundance of information, social media platforms play a crucial role as dynamic arenas for public conversation, effectively mirroring societal preoccupations, preferences, and emotions. This study sought to analyse the intricate web of public opinions and attitudes around water pollution. By utilising a range of approaches, such as wordclouds, sentiment analysis, Latent Dirichlet Allocation (LDA), and Latent Semantic Analysis (LSA), this research has effectively shed light on the intricate complexities of this significant environmental issue. The aforementioned conclusion effectively integrates the primary discoveries, emphasises the significance of each research approach, and provides valuable perspectives on the wider ramifications of the study's results.

The rapid expansion of social media has led to the transformation of platforms such as Twitter into highly valuable sources of public discourse. These platforms enable academics to explore real-time conversations, trends, and viewpoints. This study utilised a substantial corpus of user-generated content to examine the discourse surrounding water pollution in the UK. Through a rigorous examination of wordclouds and word frequencies linked to each specific locale, the study revealed a captivating assemblage of dialogues. The considerable regional disparities in the selection of pollutants, encompassing sewage, plastics, and other forms of waste, provide compelling evidence of the heterogeneous environmental concerns prevalent throughout the UK. The comprehensive examination enabled a refined classification of geographical areas according to their levels of involvement, serving as the basis for conducting sentiment analysis. The utilisation of sentiment analysis facilitated the categorisation of geographical areas based on positive and negative opinions, offering a fresh perspective to assess the extent of worries and awareness in various locations.

Among the approaches utilised, LDA and LSA have emerged as particularly enlightening. The probabilistic generative model known as LDA has exhibited its capability in comprehending the intricate levels of discourse surrounding water pollution. The six topics derived via the LDA algorithm encompassed the intricate nature of talks, encompassing a range of issues including the degradation of water quality, as well as the interconnectedness of water pollution, drinking water, and air quality. The spectrum shown in this study encompasses a wide range of conversations, effectively capturing the depth and diversity of discourse surrounding the topic. However, from the probability plot, the LDA model suggest that what people in the UK are more concerned about on the water pollution issue is their accessibility to recreational use of water. In contrast, LSA presented a perspective that was perceived as simpler and more general. Although the analysis did not delve into the specific complexities of the discourse around water contamination, it effectively emphasised the broader themes of 'water,' 'people,' and 'children.' The discrepancy observed between LDA and LSA further supports the notion that LDA's advanced and complete nature is more appropriate for capturing detailed conversations on topics, whereas LSA's simplicity may restrict its ability to capture delicate subtleties.

The implications of this research extend beyond the scope of discussions on water pollution, demonstrating the potential of Twitter data as a powerful tool for analysing society and making well-informed decisions. The results highlight the differing levels of public awareness and concern observed in different regions of the UK. This observation can function as a valuable reference for policymakers, environmental activists, and stakeholders in order to customise their actions and outreach tactics towards areas that demonstrate lower levels of knowledge and involvement.

Additionally, the sentiment analysis component provides insight into the emotional complexities present within the discussion. Regions demonstrating elevated levels of unfavourable sentiments necessitate prompt attention, indicating the pressing need to tackle prevailing issues and bridge existing knowledge disparities (Sahayak et al., 2015). In contrast, the positive sentiments shed light on

the achievements of current endeavours and establish a basis for future partnerships and endeavours aimed at raising awareness. This study provides insight into the limitless potential of utilising social media data for well-informed decision-making and sophisticated societal analysis (Otero et al., 2021). The achievement of this endeavour establishes a foundation for future investigations into further environmental issues, public health concerns, and socioeconomic complexities.

During this phase of the research, only Twitter data were employed to ascertain the geospatial data sources linked to water pollution in the United Kingdom. However, in order to enhance the depth of research, this study can be further enhanced by the utilisation of web scraping techniques to collect additional data from various sources such as web news, social media platforms, and other micro-blogging platforms. The integration of several social media platforms might enhance the ability to reach a diverse range of users and perspectives, hence expanding the scope of data collection beyond the current parameters of the proposed study. This will enhance the breadth, depth, and rigour of the analysis. Moreover, this research has the potential to be enhanced by the integration of additional sophisticated deep learning methodologies, which would facilitate the acquisition of more profound insights and a more comprehensive comprehension of the ramifications of water pollution on both the environment and human well-being. It is noteworthy to mention that this study represents the preliminary phase of a broader and more diverse investigation. The forthcoming study will entail the extraction and examination of a greater quantity of significant, comprehensive, and varied geospatial data pertaining to water pollution, facilitated by the engagement of a diverse community perspective. The incorporation of many viewpoints and beliefs from a heterogeneous cohort of persons will yield a comprehensive comprehension of the matter, hence facilitating the formulation of more efficacious resolutions and strategies to combat the issue of water pollution. The utilisation of sophisticated technologies and methodologies can effectively augment the analysis and make a valuable contribution to a more exhaustive study. This study is anticipated to yield significant insights into the matter of water pollution in the UK and make a contribution to the worldwide endeavour of attaining the net zero plus goal and a more sustainable and healthier environment.

References

- Ahuja, Shreya and Gaurav Dubey (2017). “Clustering and sentiment analysis on Twitter data”. In: *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*. IEEE, pp. 1–5.
- Albert, Alexandra, Bálint Balázs, Eglè Butkevičienė, Katja Mayer and Josep Perelló (2021). “Citizen social science: New and established approaches to participation in social research”. In: *Chapter 7. In: Vohland K. et al.(Eds). 2021. The Science of Citizen Science. Springer. <https://doi.org/10.1007/978-3-030-58278-4>. pp: 119-138.*
- Alvarez-Melis, David and Martin Saveski (2016). “Topic modeling in twitter: Aggregating tweets by conversations”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 10. 1, pp. 519–522.
- Anderson, Barbara A, John H Romani, Heston Phillips, Marie Wentzel and Kholadi Tlabela (2007). “Exploring environmental perceptions, behaviors and awareness: water and water pollution in South Africa”. In: *Population and Environment* 28, pp. 133–161.
- Barbier, Geoffrey and Liu (2011). “Data mining in social media”. In: *Social network data analytics*, pp. 327–352.
- Bashir, Ishrat, Farooq A Lone, Rouf Ahmad Bhat, Shafat A Mir, Zubair A Dar and Shakeel Ahmad Dar (2020). “Concerns and threats of contamination on aquatic ecosystems”. In: *Bioremediation and biotechnology: sustainable approaches to pollution degradation*, pp. 1–26.
- Bonney, Rick, Tina B Phillips, Heidi L Ballard and Jody W Enck (2016). “Can citizen science enhance public understanding of science?” In: *Public understanding of science* 25.1, pp. 2–16.
- Brack, Werner, Juliane Hollender, Miren López de Alda, Christin Müller, Tobias Schulze, Emma Schymanski, Jaroslav Slobodnik and Martin Krauss (2019). “High-resolution mass spectrometry to complement monitoring and track emerging chemicals and pollution trends in European water resources”. In: *Environmental Sciences Europe* 31.1, pp. 1–6.
- Brockhage, Frauke, Mientje Lüsse, Jörg Klasmeier, Verena Pietzner and Marco Beeken (Aug. 2022). “Citizen Science as an Innovative Approach to Analyze Spatial and Temporal Influences on Nitrate Pollution of Water Bodies: Results of a Participatory Research Project in Germany”. In: *Sustainability* 14, p. 9516. DOI: 10.3390/su14159516.
- Cha, YoonKyung and Craig A Stow (2015). “Mining web-based data to assess public response to environmental events”. In: *Environmental pollution* 198, pp. 97–99.
- Chandler, Mark et al. (2017). “Contribution of citizen science towards international biodiversity monitoring”. In: *Biological conservation* 213, pp. 280–294.
- Cheng, I Kit, Johannes Heyl, Nisha Lad, Gabriel Facini and Zara Grout (2021). “Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK”. In: *Scientific Reports* 11.1, p. 19009.
- Christian, Carol, Chris Lintott, Arfon Smith, Lucy Fortson and Steven Bamford (2012). “Citizen science: Contributions to astronomy research”. In: *arXiv preprint arXiv:1202.2577*.
- Cook, Sarah, Soroush Abolfathi and Nathalie I Gilbert (2021). “Goals and approaches in the use of citizen science for exploring plastic pollution in freshwater ecosystems: A review”. In: *Freshwater Science* 40.4, pp. 567–579.
- Cooper, Paul F and BC Findlater (2013). *Constructed Wetlands in Water Pollution Control: Proceedings of the International Conference on the Use of Constructed Wetlands in Water Pollution Control, Held in Cambridge, UK, 24–28 September 1990*. Elsevier.
- Cvitanic, Toni, Bumsoo Lee, Hyeon Ik Song, Katherine Fu and David Rosen (2016). “LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents”. In: *International Conference on Case-Based Reasoning*.

- Dhanwal, Pradeep, Anil Kumar, Shruti Dudeja, Vinod Chhokar and Vikas Beniwal (2017). “Recent advances in phytoremediation technology”. In: *Advances in environmental biotechnology*, pp. 227–241.
- Dittmann, Sinja, Tim Kiessling, Linda Mederake, Mandy Hinzmann, Doris Knoblauch, Marianne Böhm-Beck, Katrin Knickmeier and Martin Thiel (Aug. 2023). “Sharing communication insights of the citizen science program Plastic Pirates - best practices from 7 years of engaging schoolchildren and teachers in plastic pollution research”. In: *Frontiers in Environmental Science* 11. DOI: 10.3389/fenvs.2023.1233103.
- Edelson, Micaela, Daniel Håbesland and Rebecca Traldi (2021). “Uncertainties in global estimates of plastic waste highlight the need for monitoring frameworks”. In: *Marine Pollution Bulletin* 171, p. 112720. ISSN: 0025-326X. DOI: <https://doi.org/10.1016/j.marpolbul.2021.112720>. URL: <https://www.sciencedirect.com/science/article/pii/S0025326X21007542>.
- Feldman, Ronen (2013). “Techniques and applications for sentiment analysis”. In: *Communications of the ACM* 56.4, pp. 82–89.
- Fisher, Oliver, Nik Watson, Josep Escrig Escrig and Rachel Gomes (Jan. 2020). “Intelligent Resource Use to Deliver Waste Valorisation and Process Resilience in Manufacturing Environments”. In: 64, pp. 93–99. DOI: 10.1595/205651320X15735483214878.
- Forleo, MB and L Romagnoli (2021). “Marine plastic litter: public perceptions and opinions in Italy”. In: *Marine Pollution Bulletin* 165, p. 112160.
- Franco, Stefano and Francesco Cappa (2021). “Citizen science: involving citizens in research projects and urban planning”. In: *TeMA-Journal of Land Use, Mobility and Environment* 14.1, pp. 114–118.
- Fuller, Richard et al. (2022). “Pollution and health: a progress update”. In: *The Lancet Planetary Health* 6.6, e535–e547.
- Gacutan, Jordan, Jessica Oliver, Heidi Tait, Teerapong Praphotjanaporn and Ben Milligan (June 2023). “Exploring how citizen science projects measuring beach plastic debris can support UN Sustainable Development Goals”. In: *Citizen Science: Theory and Practice* 8. DOI: 10.5334/cstp.563.
- Gassie, Lucien W, James D Englehardt, Jian Wang, Nichole Brinkman, Jay Garland, Piero Gardinali and Tianjiao Guo (2016). “Mineralizing urban net-zero water treatment: Phase II field results and design recommendations”. In: *Water research* 105, pp. 496–506.
- Gundecha, Pritam and Liu (2012). “Mining social media: a brief introduction”. In: *New directions in informatics, optimization, logistics, and production*, pp. 1–17.
- Hasan, Md Rakibul, Maisha Maliha and M Arifuzzaman (2019). “Sentiment analysis with NLP on Twitter data”. In: *2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2)*. IEEE, pp. 1–4.
- Huddart, Joseph EA, Murray SA Thompson, Guy Woodward and Stephen J Brooks (2016). “Citizen science: from detecting pollution to evaluating ecological restoration”. In: *Wiley Interdisciplinary Reviews: Water* 3.3, pp. 287–300.
- Injadat, MohammadNoor, Fadi Salo and Ali Bou Nassif (2016). “Data mining techniques in social media: A survey”. In: *Neurocomputing* 214, pp. 654–670.
- Inyinbor Adejumo, A, O Adebisin Babatunde, P Oluyori Abimbola, A Adelani Akande Tabitha, O Dada Adewumi, A Oreofe Toyin et al. (2018). “Water pollution: effects, prevention, and climatic impact”. In: *Water Challenges of an Urbanizing World* 33, pp. 33–47.
- Jones, Oliver and Rachel Gomes (Dec. 2013). “Chemical Pollution of the Aquatic Environment by Priority Pollutants and its Control”. In: ISBN: 978-1-84973-648-0. DOI: 10.1039/BK9781849736480-00001.

- Kabir, Ahmed Imran, Ridoan Karim, Shah Newaz and Muhammad Istiaque Hossain (2018). “The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R.” In: *Informatica Economica* 22.1.
- Kalepalli, Yaswanth, Shaik Tasneem, Pasupuleti Durga Phani Teja and Suneetha Manne (2020). “Effective comparison of LDA with LSA for topic modelling”. In: *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, pp. 1245–1250.
- Kaptein, Rianne (2012). “Using Wordclouds to Navigate and Summarize Twitter Search Results.” In: *EuroHCIR*, pp. 67–70.
- Karami, Amir, Morgan Lundy, Frank Webb and Yogesh K Dwivedi (2020). “Twitter and research: A systematic literature review through text mining”. In: *IEEE access* 8, pp. 67698–67717.
- King, Abby C, Sandra J Winter, Benjamin W Chrisinger, Jenna Hua and Ann W Banchoff (2019). “Maximizing the promise of citizen science to advance health and prevent disease”. In: *Preventive medicine* 119, p. 44.
- Kousika, N and K Premalatha (2021). “An improved privacy-preserving data mining technique using singular value decomposition with three-dimensional rotation data perturbation”. In: *The Journal of Supercomputing* 77, pp. 10003–10011.
- Levin, Kelly, Taryn Fransen, Clea Schumer, Chantal Davis and Sophie Boehm (2019). “What does” net-zero emissions” mean? 8 common questions, answered”. In.
- Lin, Fan, Jianbing Xiahou and Zhuxiang Xu (2016). “TCM clinic records data mining approaches based on weighted-LDA and multi-relationship LDA model”. In: *Multimedia Tools and Applications* 75, pp. 14203–14232.
- Liu, Chloe Kliman-Silver and Alan Mislove (2014). “The tweets they are a-changin’: Evolution of twitter users and behavior”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1, pp. 305–314.
- Liu, Yihong Yuan and Fan Zhang (2020). “Mining urban perceptions from social media data”. In.
- Mavrigiannaki, A et al. (2021). “Examining the benefits and barriers for the implementation of net zero energy settlements”. In: *Energy and Buildings* 230, p. 110564.
- McGonigle, DF, RC Harris, C McCamphill, S Kirk, R Dils, J Macdonald and S Bailey (2012). “Towards a more strategic approach to research to support catchment-based policy approaches to mitigate agricultural water pollution: a UK case-study”. In: *Environmental Science & Policy* 24, pp. 4–14.
- McGregor, Shannon C (2019). “Social media as public opinion: How journalists use social media to represent public opinion”. In: *Journalism* 20.8, pp. 1070–1086.
- Mendoza, Marcelo, Bárbara Poblete and Ignacio Valderrama (2019). “Nowcasting earthquake damages with Twitter”. In: *EPJ Data Science* 8.1, pp. 1–23.
- Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela and James Rosenquist (2011). “Understanding the demographics of Twitter users”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 5. 1, pp. 554–557.
- Mohammed, Shaymaa H and Salam Al-augby (2020). “Lsa & lda topic modeling classification: Comparison study on e-books”. In: *Indonesian Journal of Electrical Engineering and Computer Science* 19.1, pp. 353–362.
- Nunavath, Vimala and Morten Goodwin (2018). “The role of artificial intelligence in social media big data analytics for disaster management-initial results of a systematic literature review”. In: *2018 5th International Conference on information and communication technologies for disaster management (ICT-DM)*. IEEE, pp. 1–4.
- O’Connor, Brendan, Ramnath Balasubramanyan, Bryan Routledge and Noah Smith (2010). “From tweets to polls: Linking text sentiment to public opinion time series”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 4. 1, pp. 122–129.

- Onan, Aytug, Serdar Korukoglu and Hasan Bulut (2016). "LDA-based topic modelling in text sentiment classification: An empirical analysis." In: *Int. J. Comput. Linguistics Appl.* 7.1, pp. 101–119.
- Otero, P, J Gago and P Quintas (2021). "Twitter data analysis to assess the interest of citizens on the impact of marine plastic pollution". In: *Marine Pollution Bulletin* 170, p. 112620.
- Pattanayak, Subhrendu K, Bruce A McCarl, Allan J Sommer, Brian C Murray, Timothy Bondelid, Dhazn Gillig and Benjamin DeAngelo (2005). "Water quality co-effects of greenhouse gas mitigation in US agriculture". In: *Climatic Change* 71, pp. 341–372.
- Periathamby, Agamuthu, SB Mehran, A Norkhairah and A Norkhairiyah (May 2019). "Marine debris: A review of impacts and global initiatives". In: *Waste Management Research* 37, p. 0734242X1984504. DOI: 10.1177/0734242X19845041.
- Plackett, Ruth, Aradhna Kaushal, Angelos P Kassianos, Aaron Cross, Douglas Lewins, Jessica Sheringham, Jo Waller and Christian von Wagner (2020). "Use of social media to promote cancer screening and early diagnosis: scoping review". In: *Journal of medical Internet research* 22.11, e21582.
- Qadri, Humaira, Rouf Ahmad Bhat, Mohammad Aneesul Mehmood and Gowhar Hamid Dar (2020). *Fresh water pollution dynamics and remediation*. Springer.
- Reyes-Garcia, Victoria, Álvaro Fernández-Llamazares, David García-del-Amo and Mar Cabeza (2020). "Operationalizing local ecological knowledge in climate change research: Challenges and opportunities of citizen science". In: *Changing Climate, Changing Worlds: Local Knowledge and the Challenges of Social and Ecological Change*, pp. 183–197.
- Rogers, David, Alun Preece, Martin Innes and Irena Spasić (2021). "Real-time text classification of user-generated content on social media: Systematic review". In: *IEEE Transactions on Computational Social Systems* 9.4, pp. 1154–1166.
- Sachdeva, Sonya and Sarah McCaffrey (2018). "Using social media to predict air pollution during California wildfires". In: *Proceedings of the 9th International Conference on Social Media and Society*, pp. 365–369.
- Sahayak, Varsha, Vijaya Shete and Apashabi Pathan (2015). "Sentiment analysis on twitter data". In: *International Journal of Innovative Research in Advanced Engineering (IJIRAE)* 2.1, pp. 178–183.
- Saldana-Perez, Magdalena, Camille Cavalière, Miguel Torres-Ruiz and Marco Moreno-Ibarra (2019). "When twitter becomes a data source for geospatial analysis". In: *Research in Computing Science* 148, pp. 357–374.
- See, Linda, Jacinto Estima, Andrea Pödör, Jamal Jokar Arsanjani, Juan-Carlos Laso Bayas and Rumiana Vatsseva (2017). "Sources of VGI for Mapping". In: *Citizen Sensor*, p. 13.
- Sen, Bulent, Mehmet Tahir Alp, Feray Sonmez, Mehmet Ali Turan Kocer and Ozgur Canpolat (2013). "Relationship of algae to water pollution and waste water treatment". In: *Water treatment* 14, pp. 335–354.
- Sorour, Shaymaa E, Tsunenori Mine, Kazumasa Godaz and Sachio Hirokawax (2014). "Comments data mining for evaluating student's performance". In: *2014 IIAI 3rd International Conference on Advanced Applied Informatics*. IEEE, pp. 25–30.
- Spielhofer, Thomas, Reynold Greenlaw, Deborah Markham and Anna Hahne (2016). "Data mining Twitter during the UK floods: Investigating the potential use of social media in emergency management". In: *2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*. IEEE, pp. 1–6.
- Stanton, Thomas, Guaduneth Chico, Elizabeth Carr, Sarah Cook, Rachel Gomes, Elizabeth Heard, Antonia Law, Hazel Wilson and Matthew Johnson (May 2022). "Planet Patrolling: A citizen science brand audit of anthropogenic litter in the context of national legislation and international

- policy". In: *Journal of Hazardous Materials* 436, p. 129118. DOI: 10.1016/j.jhazmat.2022.129118.
- Steiger, Enrico, René Westerholt, Bernd Resch and Alexander Zipf (2015). "Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data". In: *Computers, environment and urban systems* 54, pp. 255–265.
- Tavra, Marina, Ljiljana Šerić, Anka Lisec, Antonia Ivanda and Morena Galešić Divić (2021). "Transforming Social Media Posts into Volunteered Geographic Information using Data Mining Methods". In: *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)*. IEEE, pp. 1–6.
- Theobald, Ellinore J et al. (2015). "Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research". In: *Biological Conservation* 181, pp. 236–244.
- Tong, Zhou and Haiyi Zhang (2016). "A text mining research based on LDA topic modelling". In: *International conference on computer science, engineering and information technology*, pp. 201–210.
- Törnqvist, Rebecka, Jerker Jarsjö and Bakhtiyor Karimov (2011). "Health risks from large-scale water pollution: trends in Central Asia". In: *Environment international* 37.2, pp. 435–442.
- Valle-Cruz, David, Josué E Vega-Hernández and Rodrigo Sandoval-Almazán (2017). "Justice of the marquesa: A twitter trend analysis using text mining and word clouds". In: *Proceedings of the 18th Annual International Conference on Digital Government Research*, pp. 592–593.
- Vrain, Emilie and Andrew Lovett (2016). "The roles of farm advisors in the uptake of measures for the mitigation of diffuse water pollution". In: *Land Use Policy* 54, pp. 413–422.
- Wang, Huina Mao, Yuan Wang, Chris Rae and Wesley Shaw (2018). "Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data". In: *Computers & Geosciences* 111, pp. 139–147.
- Wang, Mengru, Annette BG Janssen, Jeanne Bazin, Maryna Stokal, Lin Ma and Carolien Kroeze (2022). "Accounting for interactions between Sustainable Development Goals is essential for water pollution control in China". In: *Nature communications* 13.1, p. 730.
- Wang, Xinyue Ye and Ming-Hsiang Tsou (2016). "Spatial, temporal, and content analysis of Twitter for wildfire hazards". In: *Natural Hazards* 83, pp. 523–540.
- Wang and Ligu Zhu (2017). "Research and implementation of SVD in machine learning". In: *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. IEEE, pp. 471–475.
- Wessman, Anna et al. (2019). "Citizen science in archaeology: Developing a collaborative web service for archaeological finds in Finland". In: *Transforming Heritage Practice in the 21st Century: Contributions from Community Archaeology*, pp. 337–352.
- Wiggins, Andrea and John Wilbanks (2019). "The rise of citizen science in health and biomedical research". In: *The American Journal of Bioethics* 19.8, pp. 3–14.
- Xu, Shuting, Jun Zhang, Dianwei Han and Jie Wang (2006). "Singular value decomposition based data distortion strategy for privacy protection". In: *Knowledge and Information Systems* 10, pp. 383–397.
- Yum, Seungil (2020). "Mining twitter data to understand the human sentiment on Hurricane Florence". In: *Journal of Disaster and Emergency Research*.
- Zagal, Roberto, Felix Mata and Christophe Claramunt (2016). "Geographical knowledge discovery applied to the social perception of pollution in the city of Mexico". In: *Proceedings of the 9th ACM SIGSPATIAL Workshop on Location-based Social Networks*, pp. 1–6.
- Zheng, Hang, Yang Hong, Di Long and Hua Jing (2017). "Monitoring surface water quality using social media in the context of citizen science". In: *Hydrology and Earth System Sciences* 21.2, pp. 949–961.