




Review

Twitter Data Mining for the Diagnosis of Leaks in Drinking Water Distribution Networks

Javier Jiménez-Cabas ¹, Lizeth Torres ^{2,*} and Jorge de J. Lozoya-Santos ³

¹ Departamento de Ciencias de la Computación y Electrónica, Universidad de la Costa, Barranquilla 080002, Colombia

² Instituto de Ingeniería, Universidad Nacional Autónoma de México, Ciudad de México 04510, Mexico

³ Departamento de Mecatrónica, Escuela de Ingeniería y Ciencias, Tecnológico de Monterrey, Monterrey 64849, Mexico; jorge.lozoya@tec.mx

* Correspondence: ftorreso@ingen.unam.mx

Abstract: This article presents a methodology for using data from social networks, specifically from *Twitter*, to diagnose leaks in drinking water distribution networks. The methodology involves the collection of *tweets* from citizens reporting leaks, the extraction of information from the *tweets*, and the processing of such information to run the diagnosis. To demonstrate the viability of this methodology, 358 *Twitter* leak reports were collected and analyzed in Mexico City from 1 May to 31 December 2022. From these reports, leak density and probability were calculated, which are metrics that can be used to develop forecasting algorithms, identify root causes, and program repairs. The calculated metrics were compared with those calculated through telephone reports provided by SACMEX, the entity that manages water in Mexico City. Results show that metrics obtained from *Twitter* and phone reports were highly comparable, indicating the usefulness and reliability of social media data for diagnosing leaks.

Keywords: leak diagnosis; social sensors; social network data; Twitter; text mining



Citation: Jiménez-Cabas, J.; Torres, L.; Lozoya-Santos, J.J. Twitter Data Mining for the Diagnosis of Leaks in Drinking Water Distribution Networks. *Sustainability* **2023**, *15*, 5113. <https://doi.org/10.3390/su15065113>

Academic Editor: Andreas Kanavos

Received: 16 November 2022

Revised: 1 March 2023

Accepted: 2 March 2023

Published: 14 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The world and Mexico today suffer from the water crisis that was already predicted at a meeting of the United Nations Organization (UN) held in Mar del Plata, Argentina, in 1977. More than 80% of the water in hydrological basins is already used to meet the world's demand [1]. Many regions of the world are in what is called “water stress” (the water demand is higher than the amount available) due to demographic-economic growth and climate change. Mexico, with deserts in about half of its territory, is one of these regions [2].

In the global water crisis, leaks and abuse are significant problems. For example, in the Valley of Mexico, Rio de Janeiro (Brazil), Buenos Aires (Argentina), Bucharest (Romania), Sofia (Bulgaria), and Nairobi (Kenya), approximately half of the water is wasted. Daily, around 30 million cubic meters are not invoiced due to leaks, theft, inadequate metering, and corruption [3]. The amount of water wasted could supply nearly 200 million people. The cost of this problem can be estimated at USD 141 billion a year worldwide. A third of these cases occur in developing countries, where distribution networks waste about 45 million cubic meters daily. For this reason, all possible solutions to improve the distribution and avoid the waste of drinking water are welcome.

Many approaches have been proposed to resolve the leak diagnosis problem in water distribution networks (WDN) [4], for example, methodologies based on hydraulic models, methodologies based on signals, and methodologies that take advantage of data together with hydraulic models [5–7].

In this article, we propose for the first time the mining of data from *Twitter* as an auxiliary tool for intelligent systems for the diagnosis and prognosis of leaks in water distribution networks. *Twitter* data (also known as *tweets*) is a rich source of information

on various topics (economics, politics, music, news, and events, among others). Moreover, these data can be used to find real-time trends related to a specific keyword, which can help us improve our understanding of a given situation [8].

In a fault diagnosis and prognosis context, *Twitter* users can be seen as a network of space-time sensors providing real-time information embedded in *tweets*. This information can supplement the information provided by physical sensors installed in complex systems, such as WDN, to diagnose faults. Among the most important information that a *tweet* can provide about a fault in a complex system (networks), there is the position or geolocalization, the date, the timestamp of the *tweet*, photos, and videos.

In the case of leak diagnosis, the geolocalization of a leak (seen as a fault) can have various uses, for example, to determine the density of leaks in a particular region, to help count the recurrence of some leaks, to calculate the probability of leaks in an area. It can also become a tool for users and institutions to know the exact coordinates of a leak and, based on this, provide more information on the possible causes.

Photos from *Twitter* can be used in leak diagnosis to visually assess the severity of leaks. In addition, by analyzing photos, experts can identify the type of leak and estimate the volume of water loss. This can be useful in determining the resources required to fix the fault and in prioritizing repairs. Photos can also provide additional context and information, such as the surrounding infrastructure, potential safety hazards, and the accessibility of the location. In some cases, photos can be used as evidence in legal proceedings or to hold responsible parties accountable for the damage caused by the leak. Moreover, photos can be used for training convolutional neural networks that help to classify faults.

Videos from *Twitter* can be used in fault diagnosis in several ways, such as:

1. **Visual Inspection:** Videos can provide a more detailed and dynamic view of the leak, helping experts to identify the exact location and extent of the leak. This information can help in assessing the severity of the leak and prioritizing repair work.
2. **Audio Analysis:** Videos can also capture the sound of the leak, which can help in diagnosing the type of leak and its possible cause. For example, the sound of water gushing out of a pipe may indicate a large, pressurized leak, while a faint dripping sound may indicate a smaller, slower leak.
3. **Social Context:** Videos can also provide social context to the leak, such as the presence of bystanders or the reaction of the public to the leak. This information can help in assessing the urgency of the repair work and in planning effective communication strategies.
4. **Pattern recognition:** By using machine learning algorithms, it is possible to train models on video data to recognize patterns of leak behavior.

In short, by allowing users to easily submit videos, photos, such as those shown in Figure 1, social media platforms such as *Twitter* can improve the speed and accuracy of fault reporting.

The timestamp of the *tweets* can be used to estimate repair time statistics as well as to track the evolution of the leak from the time it was first reported to the time it was fixed, even to a possible time of recurrence. By analyzing the timestamps of *tweets* reporting leaks in different locations, it is possible to identify the direction and speed of leak propagation, which can be useful in predicting the future spread of leaks and prioritizing repairs. The timestamp of a *tweet* reporting a leak can also be used to estimate the response time of the utility company or municipality in addressing the issue. This can help to identify areas where response times are particularly slow and target resources to improve efficiency.

This paper is organized as follows. Section 2 presents a summarized state of the art on using *tweets* for scientific and engineering applications. Section 3 describes how data mined from *tweets* can be used for fault diagnosis and, additionally, details a methodology to mine the location of water leaks for calculating relevant metrics for the diagnosis and prognosis of water distribution networks. Section 4 presents some results of the proposed methodology by using real data from leak reports in Mexico City. In Section 5, the advantages and

drawbacks of the proposed methodology are discussed and, finally, in Section 6, some conclusions are presented.



Figure 1. Photos shared by citizens in *Twitter* as evidence of reported leaks.

2. Related Works

Twitter is one of the most popular social networks in the world. It is estimated that it has more than 300 million users who generate 65 million daily *tweets* through their computers and mobile phones. These *tweets* share diverse information, for example, political opinions, cultural conversations, business or government propaganda, complaints, and jokes. Access to this information is an opportunity to understand trends and patterns in pandemics, natural disasters, manifestations, and accidents. An essential characteristic of the information through *Twitter* is its evolution in real time due to the restriction imposed by this social network on the length of text messages: a situation that facilitates its rapid reading and re-transmission. This restriction also encourages *Twitter* users to *tweet* multiple times a day, letting other users know what they're thinking or how everyone else is doing at the moment.

Many *tweets* result in the issuance of reports of social events such as parties, football matches, and presidential campaigns; or from disastrous events such as storms, fires, traffic jams, riots, torrential rains, earthquakes, and hurricanes. As [9] explains, each *Twitter* user can be conceived as a social sensor capable of providing space-time information on important news. That is why these social sensors have been used to conduct scientific research with social relevance. In [10], a detailed literature review is presented on mining *Twitter* data or similar short-text datasets for public health applications. According to the authors, classifying *Twitter* data into topics or categories is helpful to understand better how users react and communicate.

Table 1 provides a summary of the most recent work that utilizes *Twitter* data for monitoring, statistical analysis, and forecasting. The table presents a variety of algorithms, methods, and systems that have been proposed in recent years, which combine existing techniques with a common goal of leveraging social networks as an important source of information. The applications of these methods have ranged from natural disasters, diseases, and election campaigns to traffic status and market behavior. Notably, Table 1 highlights that the majority of recent data mining applications have focused on the social and health fields. To the best of our knowledge, there have been no reported papers that utilize these techniques for diagnosing leaks in drinking water networks. Thus,

the goal of this research is to investigate the potential use of media data from social networks, specifically *Twitter*, for detecting, locating, and evaluating leaks in drinking water distribution networks.

Table 1. *Twitter* Data Mining Recent Works.

Reference	Objective	Methodology	Application Field
[11]	To propose a generic multilayer network model for representing social networks. The model is designed to facilitate the analysis of topics in social media and can be adapted to different domains, including discussion analysis, topic analysis, and user information dissemination.	The model consists of two layers: one that describes the type of user interactions and another that connects users based on a topic extracted from their posts. The paper presents a case study on people's perception of COVID-19 vaccines by analyzing a dataset of tweets, which demonstrates the effectiveness of the proposed approach.	Health
[12]	To demonstrate the use of social network analysis to understand public discourse on <i>Twitter</i> around the novel coronavirus disease 2019 (COVID-19) pandemic. The authors examined different network properties that might affect the successful dissemination and adoption of public health messages from public health officials and health agencies.	The authors focused on conversations on <i>Twitter</i> during three key communication events from late January to early June of 2020. They employed Netlytic, a Web-based software that collects publicly available data from social media sites such as <i>Twitter</i> .	Health
[13]	To identify key characteristics of cultured meat based on communication analysis of social media.	The study used <i>Twitter</i> as a platform and analyzed 36,356 tweets posted by 4128 individual users. The analysis found that the main communicated characteristics of cultured meat were "clean meat", "future meat", and "sustainable meat". The study highlights the importance of communication in the development of the cultured meat market and the need for producers to engage with customers through appropriate communication.	Sustainability
[14]	To find out the recent major social crisis.	First, tweepplers, a trend detector online tool, is used to identify common keywords. Then, with the help of API provided by <i>Twitter</i> , a search is conducted on the recent common keywords on <i>Twitter</i> .	Social
[15]	To analyze citizens' COVID-19 vaccine hesitancy.	The authors examine three sentiment computation methods (Azure Machine Learning, VADER, and TextBlob) to analyze COVID-19 vaccine hesitancy. In addition, five learning algorithms (Random Forest, Logistics Regression, Decision Tree, LinearSVC, and Naïve Bayes) with different combinations of three vectorization methods (Doc2Vec, CountVectorizer, and TF-IDF) were deployed.	Health
[16]	To explore citizens' emotional responses in the context of the COVID-19 pandemic.	Text Network Analysis was used to identify the tweets' central issue. In addition to tweets being classified as either positive or negative, the negative sentiment was higher. The emotional responses in tweets were analyzed using the Bing and NRC (National Research Council Canada) dictionaries.	Social
[17]	To understand the United Nations World Environment Day (WED) program's impact on the digital public debate.	Tweet collection is performed by using the open-source software T-Hoarder. Subsequently, the open-source software OpenRefine and Orange Data Mining are used to prepare data for further analysis.	Social
[18]	This study aimed to understand public emotions on topics related to the COVID-19 pandemic in the United Kingdom.	Three advanced deep learning-based models are leveraged: SenticNet 6 for sentiment analysis, SpanEmo for emotion recognition, and combined topic modeling.	Social
[19]	To make a comparative text mining in 195,649 collected Persian and English tweets about JCPOA (Joint Comprehensive Plan of Action, commonly known as the Iran nuclear deal or Iran deal).	The authors present a rule-based expert system that uses the well-known concept of fingerprint in the judicial sciences to classify tweets. For detecting the unhashtagged tweets, each tweet in question checks itself with the generated fingerprint.	Social
[20]	This article presents an argument-based opinion-mining model with sentimental data analysis for extracting specific arguments.	The proposed model uses natural language processing techniques, extracting argument words to define decisions. Naïve Bayes classification is used for categorizing the results widely under agreed or disagreed.	Social
[21]	To monitor the COVID-19 disease.	This system is based on a hybridization of tweet mining and deep learning techniques.	Health
[22]	The authors focused on the potential of using <i>Twitter</i> data to extract relevant topics for the public space and investigate whether the sentiment for these topics can relate to urban design and the improvement of pedestrian space.	The research methodology included five phases: data collection (through <i>Twitter</i> API), data pre-processing (the Python library Natural Language Toolkit (NLTK) was used), data classification (conducted using the sentiment analysis tool Valence Aware Dictionary and sEntiment Reasoner-VADER), data visualization, and data analysis.	Urban Design
[23]	To detect traffic-related events and road conditions in real time.	Classification algorithms and a custom-trained named entity recognition model (NER) were used to classify and extract contextual information and visualize it on a map to get an overall picture of the traffic conditions in a city.	Traffic
[24]	To detect and locate floods in real time on a global scale.	<i>Twitter</i> 's real-time streaming (API) is used for data collection. In addition, the TAGGS [25] algorithm is used to extract locations from tweets.	Natural disasters

Table 1. Cont.

Reference	Objective	Methodology	Application Field
[26]	To identify tweets that contain signs of drug abuse and evaluate the usefulness of <i>Twitter</i> in investigating patterns of abuse over time.	Data collection and quantitative and qualitative analyses were performed manually.	Health/Social disasters
[27]	A methodology is proposed to predict crime in Chicago by combining official historical records with spatiotemporal information embedded in tweets broadcast within the area of interest.	Authors use <i>Twitter</i> -specific linguistic analysis and statistical topic modeling to identify discussion topics across an area of interest automatically.	Criminology
[9]	The authors proposed an earthquake notification system.	This system is based on a Kalman Filter and a Particle Filter. In addition, the authors designed a tweet classifier based on a support vector machine to detect a seismic event.	Natural disasters

3. Materials and Methods

Diagnosis of complex systems, such as water distribution networks, includes monitoring tasks that aim to determine the operating status of the system at all times. Fault diagnosis can be divided into three tasks [28]:

1. **Fault detection:** its goal is to identify the presence of a possible fault in the system.
2. **Fault isolation:** the purpose of this task is to pinpoint the exact cause of the fault, as well as its localization.
3. **Fault identification:** This task completes the fault diagnosis by characterizing the type of fault, its size, and its profile.

These tasks can be executed by using three types of methods: signal-based, model-based, and data-based methods [29].

1. **Signal-based methods:** The mechanism of these methods is based on the search for fault symptoms in the signals of the systems under-diagnosis, such as pressure, flow, temperature, vibrations, etc. In a healthy state, features of the acquired signals such as frequency, amplitude, phase, damping, and ripple have nominal values, which in a faulty state differ from the nominal ones. These methods can be classified according to the feature typologies: time domain techniques, frequency domain techniques and time/frequency domain techniques.
2. **Model-based methods:** These methods can be roughly classified into (1) methods without estimation error feedback and (2) methods with estimation error feedback. In the first class of methods, a mathematical model is fed with the input information of the system under diagnosis. The response of the model is compared with the response of the system. If there is a discrepancy between the responses, it is probably because there exists a fault or a set of faults. For obtaining good results with this type of method, it is necessary that the model be well calibrated in healthy conditions. Otherwise, false alarms will appear. A drawback of this class is that a disturbance can be mistaken for a fault. Another drawback of methods without estimation error feedback is that they only serve to detect faults but not to locate or isolate them. The methods with estimation feedback errors are more advantageous. The error between the model response and system response is injected into the model as an additional input and then multiplied by a gain that causes this error to approach zero over time. These methods are: (1) usefulness in detecting, locating, and isolating faults; (2) usefulness for applications in real time; (3) being robust against disturbances. These methods are also known as observer-based methods since a model with estimation error feedback is called a state observer.
3. **Data-based methods:** Methods based on models are appropriate to use when the dimension of the process (under-diagnosis) is low and it can be modeled with low-order models. However, to diagnose faults in more complex systems, these methods are no longer recommended. An alternative to diagnosing faults without using models is using data-based methods, which use information acquired from the process (under diagnosis). It can be said that these methods are the recent alternative to active supervision of systems too complex to have an explicit analytic model or

signal symptoms of faulty behavior. Data-based methods use artificial intelligence and statistics algorithms to learn from data for discovering hidden patterns in the information (system variables).

3.1. Twitter Data Mining for Fault Diagnosis

To use data-based methods to diagnose faults in complex systems, such as urban networks, sensor grids are required. In general, there are two types of grids that can be used to capture the space-time evolution of a variable, fault, disturbance, or any quantity providing information about the system behavior: (i) static sensor grids and (ii) mobile sensor grids. By analogy with the descriptions used in fluid dynamics to describe the motion of a flow, static grids can be thought of as *Eulerian Networks* because the quantities of interest are observed at fixed points (coordinates) in space. Continuing the same analogy, mobile sensor networks can be thought of as *Lagrangian Networks* because each sensor tracks and senses the quantity of interest as it moves through space and time.

Since sensors in *Eulerian Networks* have constant coordinates ($\frac{dx}{dt} = 0, \frac{dy}{dt} = 0, \frac{dz}{dt} = 0$), and the unique degree of freedom is time, they provide measurement data only dependent on time (check Figure 2a):

$$\text{Eulerian measurement: } \xi_i(t)$$

On the contrary, since sensors in *Lagrangian Networks* move, they provide measurement data that can depend on the three spatial dimensions and time (check Figure 2b):

$$\text{Lagrangian measurement: } \xi(x_i, y_i, z_i, t)$$

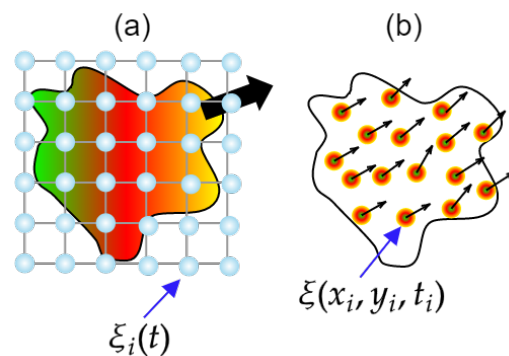


Figure 2. (a) *Eulerian Network* and (b) *Lagrangian Network*.

An example of *Eulerian Sensor Networks* is the meteorological stations installed throughout cities to provide measurements of temperature and pressure, among others. Other examples are the classical seismological systems composed of instrumented stations. Examples of *Lagrangian Sensor Networks* are the floating sensor networks, known as drifters, which take local measurements and track flows in water systems for ocean and river studies [30].

New classes of hybrid grids have emerged from classic sensor grids, incorporating a combination of fixed and mobile sensors. One example is the *Mixed Eulerian/Lagrangian Sensor Networks* (MELS Networks), which utilize both types of sensors [31]. Another example is the *Arbitrary Lagrangian-Eulerian Networks* (ALE Networks), which employ sensors that can remain inert or move in response to evolving spatial information [32]. The best and more interesting example of ALE Networks are the grids consisting of citizens equipped with smart devices, such as smartphones, tablets, or computers, and with the ability to compile and interpret what they perceive with their five senses. The components of this type of ALE Networks were called citizen-sensors by Goodchild in [33] and as social sensors in [9]. While the term ALE Network is not commonly used in the context of *Twitter* in the literature, it can be applied to describe the network of *Twitter* users who move

within a geographic area and/or those who remain in a location to report an event. In this scenario, the network topology and dynamics are subject to change over time, similar to the movement of fluid in a computational domain.

In the context of fault diagnosis, *Twitter* users can be regarded as intelligent sensors that provide information about a fault in real-time. This information is embedded in a small message with multimedia data attached (i.e., in a *tweet*) and can be extracted by using appropriate techniques.

The data extracted from the *tweet* can be used for fault diagnosis under two approaches: (i) alone or (ii) to complement the physical sensor data (quantity measurements). Under the first approach, *Twitter* users are assumed to be social sensors and can provide information about the occurrence of a fault (detection), the position of the fault (isolation), and details about the fault (identification). Under the second approach, schematized in Figure 3, the measured information provided by social sensors (y_s) is collected together with the variables acquired with physical sensors (y_m). This information is then processed to clean it and extract features from it (c). Such features are injected into artificial intelligence algorithms or statistical models to obtain patterns (p) that can be recognized to help obtain a diagnosis.

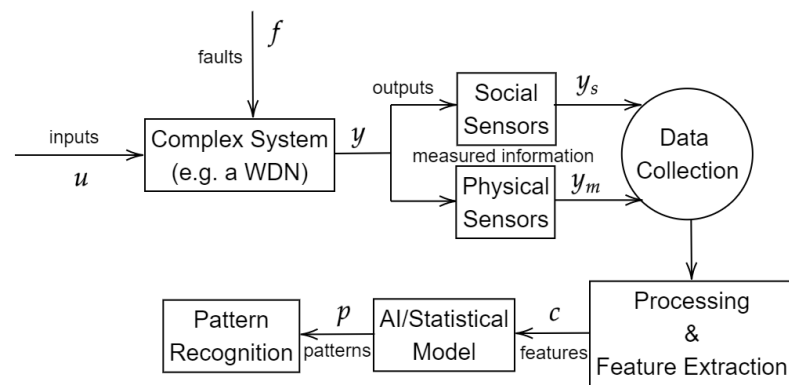


Figure 3. Data-driven methods involving social sensors.

3.2. *Twitter* Data Mining for Diagnosing Leaks

For leak diagnostics, *Twitter* users can be assumed to be position sensors that are part of a diagnostic system. However, they are not classic position sensors, they are smart position sensors that can themselves perform the three main fault diagnosis tasks: detection, localization (isolation), and identification. This last task is performed when the *Twitter* user sends pictures, videos, and more information attached to the *tweet* about the fault such as the severity of the leak, how long it has been active, the damages as well as the consequences that the leak is provoking; see Figure 4. The information from the diagnosis can then be used for fault-tolerant control, for calculating the fault rate, and for prognosis algorithms.

Following, a methodology to obtain the geographical coordinates of water leaks reported by *Twitter* users is presented. This methodology is inspired by the series of steps proposed by [34] to mine *tweet* content. The six steps that make up this methodology are illustrated in Figure 5 and described below. To describe each of the steps, the example of extracting leak positions in Mexico City is used. However, this methodology can be generalized and used for the diagnosis of leaks in other cities or conurbation areas of the world.

3.2.1. Collection of Tweets

The first step is the collection of *tweets* according to specific search rules and restrictions. The collection consists of two stages: the search and storage of *tweets*, tasks that can be performed by using an application, either commercial or self-authored. To collect *tweets* in which leaks are reported in Mexico City, *tweets* in Spanish that include any of the following words in the body of the message are searched for: {fuga}, {fugando}, or {SACMEX} (Agency

that manages water in Mexico City). Given that leaks of gas, fuel, information, or criminals are also reported, it is necessary to exclude *tweets* containing the words {gas, combustible} or phrases such as {se dió a la fuga, fuga de divisas}. In this way, there are three sets of words: (1) those that are searched for in the *tweet*, (2) those that the *tweet* should not contain, and (3) those that are associated with a specific structure, which should not be in the *tweet* either. Figure 6 shows some leak reports found from the execution of the search rules.

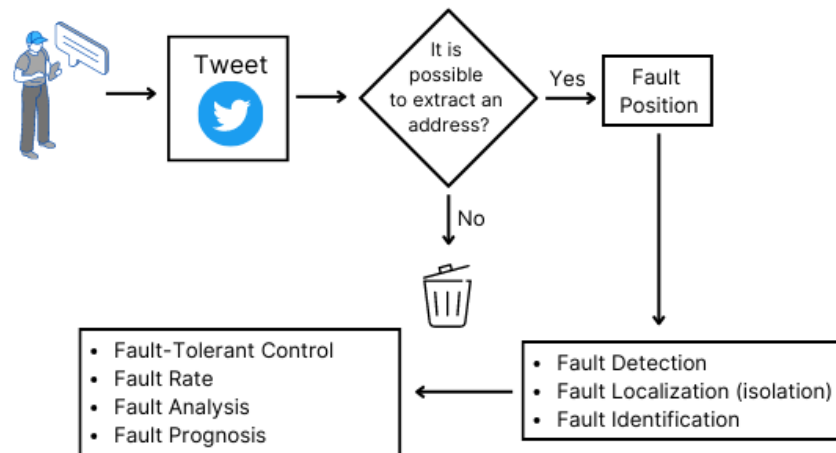


Figure 4. Approach for Content Mining of *Tweets*.

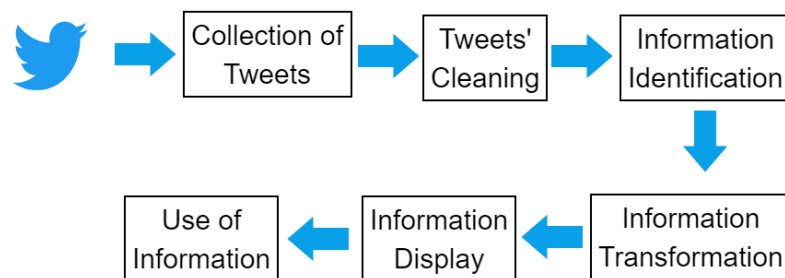


Figure 5. Approach for diagnosing leaks using social sensors.



Figure 6. Screenshots of *tweets* of Citizens Reporting Leaks.

Tweets that meet the search rules are downloaded into spreadsheets containing the *tweet* metadata and the user’s profile information as date, username, *tweet* text, *tweet* ID, location, media files attached to the *tweet*, among others. Figure 7 presents the metadata and profile information of the *tweets* that were downloaded for this work.

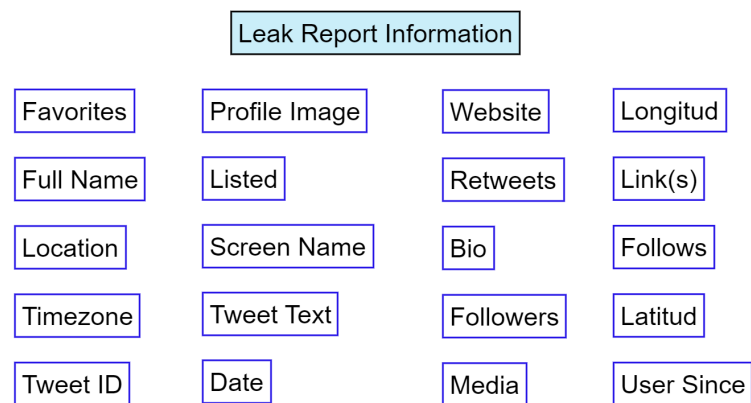


Figure 7. Metadata and user's profile information useful for leak diagnosis.

3.2.2. Tweets Cleaning

The text cleaning process or noise filtering, within the scope of text mining, consists of removing from the text everything that does not provide information about its theme, structure, or content such as characters, words, and *emojis*. Therefore, a series of filters must be applied, such as those proposed by [34,35]. There is no single way to do it, it largely depends on the purpose of the analysis and the source from which the text comes. While some writing guides recommend removing conjunctions or prepositions, it is important to note that in the case of addresses in Mexico City, the conjunction "y" (which means "and" in Spanish) should not be deleted. This is because it is often used to indicate the intersection or union of two streets. Therefore, it is crucial to include it in the address to avoid confusion or misinterpretation of the location.

3.2.3. Information Identification

In this step, the desired information is extracted, for example, names of people, books, movies or postal codes. For leak diagnosis, the task to accomplish in this stage is the identification of postal addresses in the text of a *tweet*. Then, an application can be chosen or developed that uses a particular identification strategy. In the case of postal addresses, the identification process varies from country to country [36].

In Mexico, some keywords allow us to identify a postal address, for example, {Avenida} or its abbreviation {Av.}, {Calle} or its abbreviation {C.}, the abbreviation of number {No.}, the symbol {#}, {Esquina}, and its abbreviation {Esq.} For identification, arrays of numbers from 1 to 4 digits are also sought.

3.2.4. Transformation of Information

In this step, the information is converted into numerical values with a meaning in the context in which it will be used. For leak diagnosis, the task to be executed in this stage is the geocoding of postal addresses, that is, the transformation of postal addresses to latitude and longitude coordinates. This task can be performed manually or automatically with the help of an application. To accomplish this task, we developed an application on the Google Apps Script platform. We integrated it into the spreadsheets in which the *tweets* that satisfy the search rules are stored so that the coordinates (latitude and longitude) are automatically calculated from the postal address contained in each *tweet*.

3.2.5. Information Display

The goal of this step is to present the information in a meaningful and visual way for its interpretation and understanding. Information visualization is an effective way to share knowledge in a digestible format that helps make the best use of it. To effectively display the information of the coordinates of the leaks, the *Google My Maps* service was used. For this, an application was developed on the *Google Apps Script* platform, which automatically places a marker at each coordinate where a water leak was reported.

The methodology up to this step is presented in Algorithm 1 for ease of understanding.

Algorithm 1 *Twitter Search and Filtering Algorithm*

- 1: Set the sets of search words and spam words denoted by Q and E , respectively.
 - 2: Set the search radius r in kilometers.
 - 3: Set the time interval t in minutes for searching *Twitter*.
 - 4: Set the maximum number of tweets to retrieve n .
 - 5: Set the time period P for searching *Twitter*.
 - 6: Set the output file name F for saving the obtained tweets and their geographic coordinates.
 - 7: Set the API keys for accessing the *Twitter* API.
 - 8: Initialize an empty set of tweets T .
 - 9: Initialize an empty set of spam tweets S .
 - 10: **while** T contains less than n tweets and T has not been updated for P minutes **do**
 - 11: Search *Twitter* API for tweets containing any word in Q within a radius of r kilometers from a central location.
 - 12: Add any new tweets to T .
 - 13: Discard any tweets from T that contain one or more words contained in E .
 - 14: Add any discarded tweets to S .
 - 15: Sleep for t minutes.
 - 16: **end while**
 - 17: For each tweet $t_w \in T$, obtain its geographic coordinates (latitude, longitude).
 - 18: Save the set of tweets T and their geographic coordinates to a file F .
 - 19: Place an indicator on a map at each coordinate obtained in the previous step.
 - 20: **return** T , S , and the map with indicators.
-

3.2.6. Use of Information

The mined information can be used to plan, make decisions, schedule activities, create strategies and compute metrics or performance indexes that can be used to improve the performance of WDN. Some of these metrics, which are usually employed by the organizations that manage water, are defined below.

- **Leak density** is a metric used to describe the frequency or concentration of leaks in a water distribution network. It is typically measured as the number of leaks per unit area (e.g., leaks per square kilometer). The calculation of leak density involves dividing the total number of leaks in a given area by the size of that area. The leak density can be used by water distribution companies to prioritize maintenance or repair work in the areas with the highest leak density. By addressing leaks in these areas, water utilities can reduce water losses, improve the efficiency of the distribution system, and potentially save money on water treatment and pumping costs. Additionally, the leak density can be used to identify potential areas of pipe corrosion or damage, or to detect patterns in the occurrence of leaks, which can help inform decisions on pipe replacement or rehabilitation.
- **Leak frequency** in a water distribution network can be calculated by dividing the total number of leaks that occurred in a given period by the total length of the water distribution network. The resulting value indicates the average number of leaks that occurred per unit length of the network during that period. For example, if there were 10 leaks in a water distribution network of 10 kilometers in a month, the leak frequency would be 1 leak per kilometer per month (10 leaks \div 10 kilometers = 1 leak/kilometer). By tracking the frequency of leaks over time, water utility companies can identify areas of the network that are prone to leaks and prioritize maintenance and repair efforts in those areas.
- **Leak probability** in an area refers to the likelihood or chance of a leak occurring in a specific region of a water distribution network. It is usually calculated based on historical data about leak occurrences in the area and other relevant factors such as

the age and condition of the pipes, water pressure, and environmental conditions. By estimating the leak probability, water utilities and managers can prioritize maintenance and repair efforts, allocate resources more effectively, and proactively address potential leaks before they become significant issues.

In conclusion, accomplishing this step is the goal of *tweet* mining for leak diagnosis. The methodology up to this step is presented in Algorithm 2 facilitate its understanding.

Algorithm 2 *Twitter*-based Leak Diagnosis Algorithm

- 1: Set the sets of search words and spam words denoted by Q and E , respectively.
 - 2: Search *Twitter* every t minutes for words included in the set Q and obtain the set of *tweets* T that contain them.
 - 3: Given a set E , discard *tweets* from set T that contain one or more words contained in set E .
 - 4: For each $t_{wi} \in T$, obtain its geographic coordinates (latitude, longitude).
 - 5: For each municipality M_i in the city:
 - 6: Compute the number of *tweets* n_{wi} containing words in set Q that are geographically located within M_i .
 - 7: Compute the total number of *tweets* N containing words in set Q that are geographically located within the area composed by all the municipalities.
 - 8: Compute the density of leaks for M_i as $D_i = n_{wi} / A_i$, where A_i is the area of M_i .
 - 9: Compute the average density of leaks of all the municipalities as $\bar{D} = \frac{1}{N} \sum_{i=1}^N D_i$.
 - 10: Compute the probability of leaks for M_i as $P_i = D_{wi} / \bar{D}$.
 - 11: Compute the percentage of leak reports for M_i as $L_i = n_{wi} / N$.
 - 12: 6. Return the sets of densities D , probabilities P , and percentages L for all municipalities.
-

4. Results

The data presented below were obtained from 1 May 2022 to 31 December 2022. During this period, 358 *tweets* were found that satisfied the search conditions for Mexico City. In the same period, 8508 reports of leaks were registered via telephone by the water agency (SACMEX). Figure 8 displays the percentages of the reports of leaks via telephone and via *Twitter* for each municipality. These percentages were calculated as follows:

$$\% \text{ of } Twitter \text{ Reports in a Municipality} = \frac{\text{Twitter Reports in the Municipality}}{\text{Total of Twitter Reports in Mexico City}} \times 100$$

It can be seen that the municipalities with the most reports via telephone are Gustavo A. Madero, Iztapalapa, and Tlalpan, while the municipalities with the most reports via *Twitter* are Coyoacán, Iztapala, and Tlalpan. This means that Iztapala and Tlalpan are the municipalities that report leaks the most, both on *Twitter* and by phone.

The density and probability of leaks were calculated from reports received via *Twitter* and phone. As shown in Table 2, the three municipalities with the highest density and probability of leaks based on phone reports were Benito Juárez, Cuauhtémoc, and Gustavo A. Madero. Meanwhile, based on *Twitter* reports, the municipalities with the highest density and probability of leaks were Benito Juárez, Coyoacán, and Cuauhtémoc.

Notice that the results obtained from calculating leak probability using both phone and *Twitter* reports are somewhat similar. This suggests that data from social media can be a reliable and useful source of information for identifying and diagnosing leaks in water distribution networks. This also indicates that the proposed methodology of using social media data can be a viable alternative to traditional methods of collecting leak reports through phone calls or other means.

Figure 9 shows the Mexico City map with red marks indicating the positions of the reported leaks on *Twitter*. In addition, the map includes the geographical limits of

the municipalities in Mexico City, which allows us to visualize the density of leaks for each region.

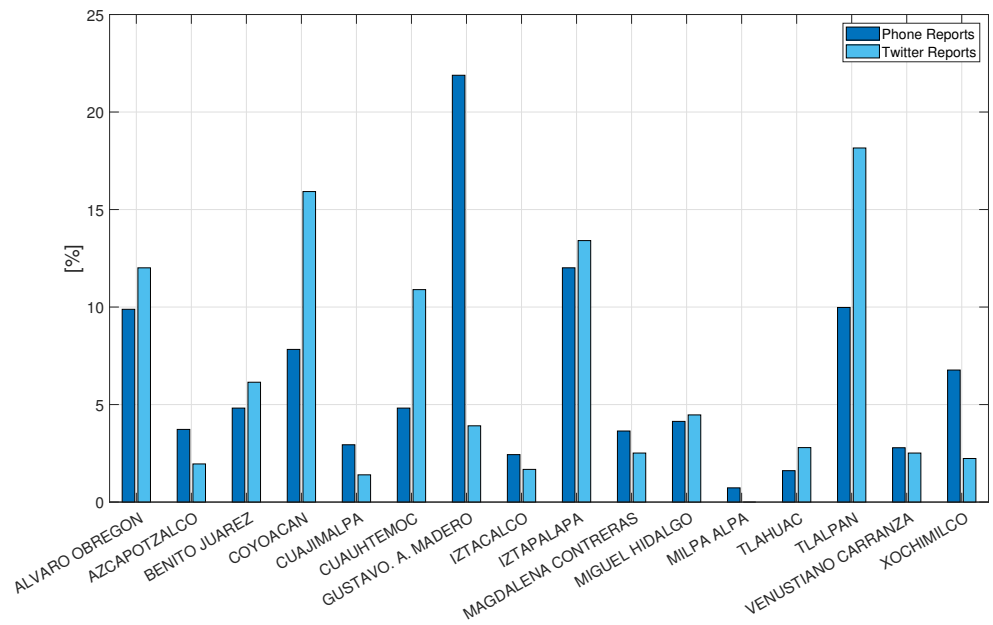


Figure 8. Percentages of leak reports for each municipality from 1 May 2022, to 31 December 2022. It is important to note that the percentage of leaks reported by telephone was calculated based on the total number of reports made solely through that channel. Likewise, the percentage of leak reports by municipality on Twitter was calculated from the total number of reports made on Twitter city-wide.

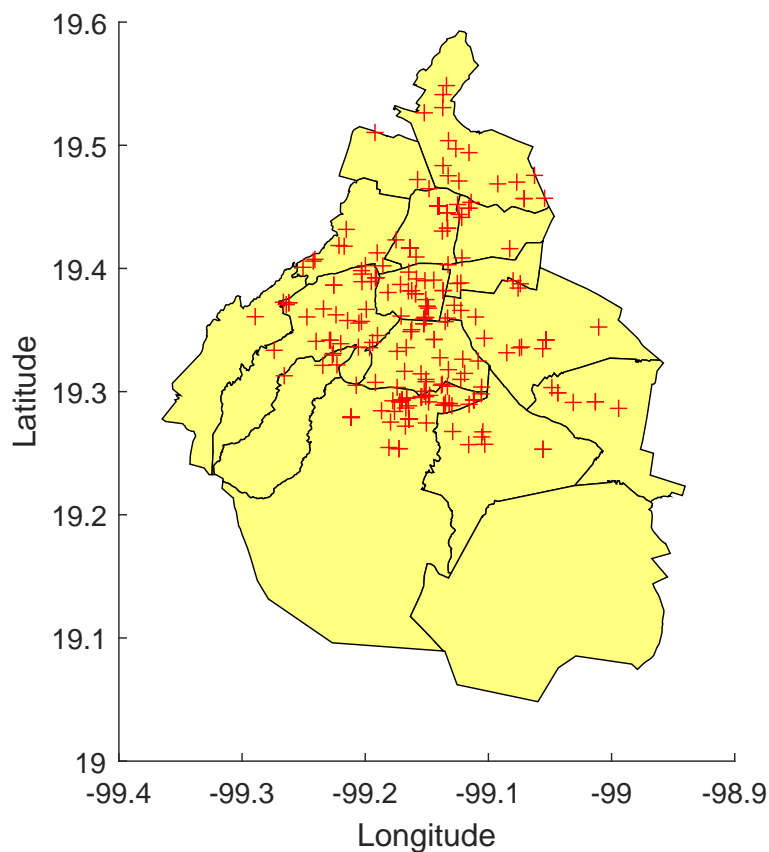


Figure 9. Map of Mexico City displaying the locations of the reported leaks on *Twitter*.

Table 2. Leak density and probability calculated from *tweets* and phone reports.

Municipality	Phone Reports (PR)	Twitter Reports (TR)	Surface (km ²)	Density PR/km ²	Density TR/km ²	Probability from PR	Probability from TR
ALVARO OBREGON	841	43	96.17	8.744930852	0.447124883	1.10	1.25
AZCAPOTZALCO	317	7	33.66	9.417706477	0.207961973	1.19	0.58
BENITO JUAREZ	410	22	26.63	15.39616973	0.826135937	1.94	2.31
COYOACAN	666	57	54.40	12.24264706	1.047794118	1.54	2.92
CUAJIMALPA	250	5	80.95	3.088326127	0.061766523	0.39	0.17
CUAUHTEMOC	410	39	32.40	12.65432099	1.203703704	1.59	3.36
GUSTAVO. A. MADERO	1862	14	94.07	19.7937706	0.148825343	2.49	0.42
IZTACALCO	207	6	23.30	8.884120172	0.25751073	1.12	0.72
IZTAPALAPA	1022	48	117.00	8.735042735	0.41025641	1.10	1.15
MAGDALENA CONTRERAS	310	9	74.58	4.156610351	0.120675784	0.52	0.34
MIGUEL HIDALGO	352	16	46.99	7.490955522	0.340497978	0.94	0.95
MILPA ALPA	62	0	228.41	0.271441706	0	0.03	0.00
TLAHUAC	137	10	85.34	1.605343333	0.117178345	0.20	0.33
TLALPAN	849	65	312.00	2.721153846	0.208333333	0.34	0.58
VENUSTIANO CARRANZA	237	9	33.40	7.095808383	0.269461078	0.89	0.75
XOCHIMILCO	576	8	122.00	4.721311475	0.06557377	0.59	0.18

5. Discussion

There are several advantages to using leak reports made using *Twitter* over leak reports via telephone for calculating metrics:

1. **Real-time reporting:** *Twitter* allows for instant reporting of leaks, while reporting leaks via telephone may take longer as individuals may have to wait on hold or navigate automated systems.
2. **More detailed information:** *Twitter* allows for more detailed information to be provided, including pictures or videos of the leaks, which can aid in identifying the location and severity of the leak.
3. **Cost-effective:** *Twitter* reporting is a cost-effective method of identifying leaks, as it does not require the same amount of resources as traditional telephone reporting systems.

Some disadvantages of using *Twitter* reports are:

1. **Limited coverage:** Not everyone uses *Twitter* or social media, so the sample of reports may not be representative of the entire population. This can lead to underreporting or biased reporting of leaks.
2. **Data privacy:** *Twitter* reports may contain personal information or sensitive data that may need to be protected. Water utility companies may need to ensure that the data collected from social media platforms are compliant with data protection regulations.

The next future works of the research presented in this article are two: the realization of a sensitivity analysis to test the impact of the increase of important parameters in the results of the algorithms, as well as the use of videos and photos to classify the severity of leaks and give them a prioritizing weight for their repair.

For sensitivity analysis, different versions of the algorithms presented in this article will be run simultaneously. Each of these versions will have a modified parameter, for example, the search radius or keywords. The results obtained by the different versions of the algorithms will be analyzed to determine how the variations of the modified parameters affect the results. This analysis can help to refine and improve the method for more accurate and reliable leak diagnosis in drinking water distribution networks based on tweet reports.

For the classification of the leaks by severity with photos, the approach to be used is computer vision using deep learning techniques. Deep learning models can be trained to recognize and classify different types of leaks in images and videos, and can be used to automatically determine the severity of a leak based on factors such as the size, location, and shape of the leak.

One example of a deep learning model that can be used for leak severity classification is a convolutional neural network (CNN). CNNs are designed to recognize patterns in images and can be trained to identify specific features of leaks that are associated with different levels of severity. For example, a CNN could be trained to recognize the difference between a small leak and a large leak, or between a leak that is easy to repair and one that requires more extensive repairs.

6. Conclusions

This article is the first result of a research work with the primary motivation to demonstrate that social networks are an excellent alternative for fault detection in large city networks using citizens as intelligent sensors. Hence, a new archetype for diagnosing faults in complex systems based on mining data from *Twitter* was presented. This archetype is a class of data-based methods in which the data are acquired by intelligent sensors (citizen using *Twitter*) that can send information about a fault, such as its position, size, severity, etc. The method was tested and applied to extract information related to the geographical location of leaks in the different municipalities of Mexico City. The mined information can feed forecast models and warning systems, create historical records for statistical analysis and risk assessment, and make leak events visible for their immediate attention.

Author Contributions: Conceptualization, L.T.; methodology, L.T. and J.J.-C.; validation, J.J.-C. and J.d.J.L.-S.; writing original draft, J.J.-C.; writing, review and editing, L.T. and J.d.J.L.-S. All authors discussed the results and have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: Data available on request due to privacy restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UN	United Nations
WDN	Water Distribution Networks
GIS	Geographic Information System
API	Application Programming Interface
NRC	National Research Council Canada
JMA	Japan Meteorological Agency
MELSN	Mixed Eulerian/Lagrangian Sensor Network
ALEN	Arbitrary Lagrangian-Eulerian Network
JCPOA	Joint Comprehensive Plan of Action
NLTK	Natural Language Toolkit
VADER	Valence Aware Dictionary and Sentiment Reasoner
NER	Named Entity Recognition Model
CNN	Convolutional Neural Network
MELSN	Mixed Eulerian/Lagrangian Sensor Network
ALEN	Arbitrary Lagrangian-Eulerian Network
AI	Artificial Intelligence

References

1. Ling, T. A Global Study about Water Crisis. In Proceedings of the 2021 International Conference on Social Development and Media Communication (SDMC 2021), Sanya, China, 26–28 November 2021; Atlantis Press: Paris, France, 2022; pp. 809–814.
2. Briseño, H.; Sánchez, A. Decentralization, consolidation, and crisis of urban water management in Mexico. *Tecnol. y Cienc. Del Agua* **2018**, *9*, 25–47. [[CrossRef](#)]

3. Khalifa, D.S.; El Atty, A.; Donia, N.S.; Moussa, A.; Mohamed, A. Analysis and Assessment of Water Losses in Domestic Water Distribution Networks. *J. Environ. Sci.* **2022**, *51*, 1–23. [[CrossRef](#)]
4. Verde, C.; Torres, L. *Modeling and Monitoring of Pipelines and Networks: Advanced Tools for Automatic Monitoring and Supervision of Pipelines*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 7.
5. Carpentier, P.; Cohen, G. State estimation and leak detection in water distribution networks. *Civ. Eng. Syst.* **1991**, *8*, 247–257. [[CrossRef](#)]
6. Pérez, R.; Puig, V.; Pascual, J.; Quevedo, J.; Landeros, E.; Peralta, A. Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks. *Control Eng. Pract.* **2011**, *19*, 1157–1167. [[CrossRef](#)]
7. Soldevila, A.; Blesa, J.; Tornil-Sin, S.; Duviella, E.; Fernandez-Canti, R.M.; Puig, V. Leak localization in water distribution networks using a mixed model-based/data-driven approach. *Control Eng. Pract.* **2016**, *55*, 162–173. [[CrossRef](#)]
8. Li, X.; Wen, Y.; Jiang, J.; Daim, T.; Huang, L. Identifying potential breakthrough research: A machine learning method using scientific papers and Twitter data. *Technol. Forecast. Soc. Chang.* **2022**, *184*, 122042. [[CrossRef](#)]
9. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 851–860.
10. Jordan, S.E.; Hovet, S.E.; Fung, I.C.H.; Liang, H.; Fu, K.W.; Tse, Z.T.H. Using Twitter for public health surveillance from monitoring and prediction to public response. *Data* **2018**, *4*, 6. [[CrossRef](#)]
11. Bonifazi, G.; Breve, B.; Cirillo, S.; Corradini, E.; Virgili, L. Investigating the COVID-19 vaccine discussions on Twitter through a multilayer network-based approach. *Inf. Process. Manag.* **2022**, *59*, 103095. [[CrossRef](#)]
12. Pascual-Ferrá, P.; Alperstein, N.; Barnett, D.J. Social Network Analysis of COVID-19 Public Discourse on Twitter: Implications for Risk Communication. *Disaster Med. Public Health Prep.* **2022**, *16*, 561–569. [[CrossRef](#)] [[PubMed](#)]
13. Pilařová, L.; Kvasničková Stanislavská, L.; Pilař, L.; Balcarová, T.; Pitrová, J. Cultured Meat on the Social Network Twitter: Clean, Future and Sustainable Meats. *Foods* **2022**, *11*, 2695. [[CrossRef](#)] [[PubMed](#)]
14. Rahman, S.; Jahan, N.; Sadia, F.; Mahmud, I. Social crisis detection using Twitter based text mining—a machine learning approach. *Bull. Electr. Eng. Inform.* **2023**, *12*, 1069–1077. [[CrossRef](#)]
15. Qorib, M.; Oladunni, T.; Denis, M.; Ososanya, E.; Cotae, P. COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination twitter dataset. *Expert Syst. Appl.* **2023**, *212*, 118715. [[CrossRef](#)]
16. Choi, Y.J.C.E.J. The Early Emotional Responses and Central Issues of People in the Epicenter of the COVID-19 Pandemic: An Analysis from Twitter Text Mining. *Int. J. Ment. Health Promot.* **2023**, *25*, 21–29. [[CrossRef](#)]
17. Zarrabeitia-Bilbao, E.; Rio-Belver, R.M.; Alvarez-Meaza, I.; de Alegria-Mancisidor, I.M. World Environment Day: Understanding Environmental Programs Impact on Society Using Twitter Data Mining. *Soc. Indic. Res.* **2022**, *164*, 263–284. [[CrossRef](#)]
18. Alhuzali, H.; Zhang, T.; Ananiadou, S. Emotions and Topics Expressed on Twitter During the COVID-19 Pandemic in the United Kingdom: Comparative Geolocation and Text Mining Analysis. *J. Med Internet Res.* **2022**, *24*, e40323. [[CrossRef](#)]
19. Behzadidoost, R.; Hasheminezhad, M.; Farshi, M.; Derhami, V.; Alamiyan-Harandi, F. A framework for text mining on Twitter: A case study on joint comprehensive plan of action (JCPOA)-between 2015 and 2019. *Qual. Quant.* **2022**, *56*, 3053–3084. [[CrossRef](#)]
20. Arumugam, S.S. Development of argument based opinion mining model with sentimental data analysis from twitter content. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6956. [[CrossRef](#)]
21. Jiang, J.Y.; Zhou, Y.; Chen, X.; Jhou, Y.R.; Zhao, L.; Liu, S.; Yang, P.C.; Ahmar, J.; Wang, W. COVID-19 Surveiller: Toward a robust and effective pandemic surveillance system based on social media mining. *Philos. Trans. R. Soc. A* **2022**, *380*, 20210125. [[CrossRef](#)]
22. Vukmirovic, M.; Raspopovic Milic, M.; Jovic, J. Twitter Data Mining to Map Pedestrian Experience of Open Spaces. *Appl. Sci.* **2022**, *12*, 4143. [[CrossRef](#)]
23. Khetarpaul, S.; Sharma, D.; Jose, J.I.; Saragur, M. Real-Time Detection and Visualization of Traffic Conditions by Mining Twitter Data. In Proceedings of the Australasian Database Conference, Sydney, Australia, 3–4 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 141–152.
24. de Bruijn, J.A.; de Moel, H.; Jongman, B.; de Ruyter, M.C.; Wagemaker, J.; Aerts, J.C. A global database of historic and real-time flood events based on social media. *Sci. Data* **2019**, *6*, 1–12. [[CrossRef](#)]
25. De Bruijn, J.A.; de Moel, H.; Jongman, B.; Wagemaker, J.; Aerts, J.C. TAGGS: Grouping tweets to improve global geoparsing for disaster response. *J. Geovisualization Spat. Anal.* **2018**, *2*, 1–14. [[CrossRef](#)]
26. Sarker, A.; O’connor, K.; Ginn, R.; Scotch, M.; Smith, K.; Malone, D.; Gonzalez, G. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. *Drug Saf.* **2016**, *39*, 231–240. [[CrossRef](#)] [[PubMed](#)]
27. Gerber, M.S. Predicting crime using Twitter and kernel density estimation. *Decis. Support Syst.* **2014**, *61*, 115–125. [[CrossRef](#)]
28. Isermann, R. *Fault-Diagnosis Applications: Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors, and Fault-Tolerant Systems*; Springer Science & Business Media: Berlin, Germany, 2011.
29. Gonzalez-Jimenez, D.; Del-Olmo, J.; Poza, J.; Garramiola, F.; Madina, P. Data-driven fault diagnosis for electric drives: A review. *Sensors* **2021**, *21*, 4024. [[CrossRef](#)]
30. Tinka, A.; Rafiee, M.; Bayen, A.M. Floating sensor networks for river studies. *IEEE Syst. J.* **2012**, *7*, 36–49. [[CrossRef](#)]
31. Canepa, E.; Odat, E.; Dehwah, A.; Mousa, M.; Jiang, J.; Claudel, C. A sensor network architecture for urban traffic state estimation with mixed eulerian/lagrangian sensing based on distributed computing. In Proceedings of the International Conference on Architecture of Computing Systems, Lubeck, Germany, 25–28 February 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 147–158.

32. Hirt, C.; Amsden, A.; Cook, J. An arbitrary Lagrangian-Eulerian computing method for all flow speeds. *J. Comput. Phys.* **1974**, *14*, 227–253. [[CrossRef](#)]
33. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [[CrossRef](#)]
34. Yoon, S.; Elhadad, N.; Bakken, S. A practical approach for content mining of tweets. *Am. J. Prev. Med.* **2013**, *45*, 122–129. [[CrossRef](#)]
35. Ralston, M.R.; O'Neill, S.; Wigmore, S.J.; Harrison, E.M. An exploration of the use of social media by surgical colleges. *Int. J. Surg.* **2014**, *12*, 1420–1427. [[CrossRef](#)]
36. Kayed, M.; Dakrory, S.; Ali, A.A. Postal address extraction from the web: A comprehensive survey. *Artif. Intell. Rev.* **2021**, *55*, 1085–1120. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.