Computing&AI Connect
ISSN: 3104-4719
2024, Article ID. 2024.0006, Cite as: https://doi.org/10.69709/CAIC.2024.133109

SCIFINITI PUBLISHING

Research Article

OPEN ACCESS

Social Media and Artificial Intelligence for Sustainable Cities and Societies: A Water Quality Analysis Use-Case

Muhammad Asif Ayub[✉,1] **Muhammad Tayyab Zamir**[✉,2] **Imran Khan**[✉,1] **Hannia Naseem**[✉,1] **Nasir Ahmad**[✉,3] **Kashif Ahmad**[✉,3,*]

[1] Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan
[2] Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico
[3] Department of Computer Science, Munster Technological University, Cork, Ireland

Article History
Received: September 08, 2024 Accepted: November 06, 2024 Published: December 12, 2024

Abstract
Crowd-sourcing has been widely explored for monitoring and feedback on infrastructure and services, such as air and water quality analysis. However, the traditional methods of crowd-sourcing for feedback and analysis of water quality, namely offline and online surveys, have several limitations, such as the limited number of participants and low frequency due to the labor involved in conducting such surveys. Social media analytics could overcome these challenges by providing a more sustainable and cost-effective water quality monitoring and analysis tool. This paper explores the potential of social media analytics in addressing water quality issues by proposing a Natural Language Processing (NLP) framework that automatically collects and analyses water-related posts to support data-driven decision-making. The proposed framework is composed of two components, namely (i) text classification, and (ii) topic modeling. The study proposes a merit-fusion-based framework for text classification, incorporating several Large Language Models (LLMs) combined in a late fusion method with optimal weights. In topic modeling, the BERTopic library was employed to discover the hidden topic patterns in the water-related tweets. Relevant tweets were also analyzed, originating from different regions and countries to explore global, regional, and country-specific issues and water-related concerns. A large-scale dataset was also collected and manually annotated, which is expected to facilitate future research on the topic.

Keywords:
smart cities; sustainable cities; Artificial Intelligence (AI); Natural Language Processing (NLP); water quality; topic modelling; named entity recognition; text classification; Large Language Models (LLMs)

1. Introduction

Real-time monitoring and observation of resources and infrastructure are primary tasks for achieving resilient infrastructure and sustainable cities [1]. This allows for appropriate recovery actions to mitigate risks and damages. The literature reports several interesting methods for monitoring infrastructure [2]. Crowd-sourcing is one of the methods enabling real-time monitoring and feedback on infrastructure [3]. Crowd-sourcing could be carried out in several ways. For instance, one of the key and widely explored crowd-sourcing methods is conducting surveys for citizens' feedback on different services, such as water quality, air quality, roads, infrastructure, and other societal challenges. These surveys can help in obtaining more detailed, contextual, and localized information [4]. These surveys are either conducted by asking citizens to fill in an online form or a questionnaire. More recently, mobile applications have also been developed for conducting such surveys, where the participants are asked to install the application and give feedback. However, these online and in-person surveys have several limitations as well [5]. One of the key limitations of such surveys is the limited scope, which means they can cover a limited number of

* **Corresponding Author:**
Kashif Ahmad, Department of Computer Science, Munster Technological University, Cork, Ireland, kashif.ahmad@mtu.ie

participants as people are often reluctant to install such applications, as noticed during the COVID-19 pandemic [6]. Moreover, conducting surveys is time-consuming and requires significant human and other resources. As a result, the frequency of such surveys is generally low as it is costly and unfeasible to frequently conduct such surveys.

These limitations of the current crowd-sourcing methods could be overcome by extracting information from social media outlets, such as Twitter and Facebook. Social media outlets have already been proven to be an effective source of communication and information spreading [7,8]. Their capabilities to engage large volumes of audiences worldwide make them a preferred platform for discussing and conveying concerns over different domestic and global challenges. The literature already reports their effectiveness in a diversified set of societal, environmental, and technological topics [9,10].

In this work, the potential of social media has been explored as a crowd-sourcing source/medium of instant feedback on water quality. To this aim, an automatic solution is proposed allowing the collection and analysis of citizens' feedback on water quality. The proposed system will not only engage large participants, which is a key factor in meaningful feedback, but it will also be a continuous process and will keep collecting people's feedback continuously. One of the key advantages of the system is its ability to collect and analyze feedback without requiring citizens to fill out online forms or surveys. Instead, it continuously filters and analyzes relevant social media posts in a privacy-preserving manner, anonymizing personal information that could identify individuals.

The proposed system is composed of (i) a crawler, which is responsible for collecting social media posts (Tweets), (ii) a classification framework employing several NLP algorithms in a merit-based fusion to differentiate between water-related and irrelevant tweets, and (iii) topical analysis to automatically analyze and extract key water-related issues discussed in the tweets. For the training and evaluation of the text classification framework, a large-scale benchmark dataset was collected and annotated, which will be made publicly available for further research in the domain.

The key contributions of the work can be summarized as follows:

- Proposing an automatic tool to collect, analyze, and extract meaningful information from social media posts as a source of instant feedback on water quality, as a first step towards a sustainable water network.
- Proposing a merit-based fusion framework by combining several transformers-based NLP algorithms to differentiate between water-related and irrelevant tweets.
- Collecting and annotating a large-scale benchmark dataset containing around 8,000 tweets.
- Performing topic modeling on the relevant tweets to automatically extract key water-related issues discussed in the relevant tweets.
- Analyzing the origin of the water-related tweets and providing region and country-wise distribution of the water-related tweets collected by our system. This analysis shows the growing concern over this important societal challenge.

The rest of the paper is organized as follows. Section 2 provides an overview of the related work. Section 3 discusses the proposed methodology. Section 4 covers the experimental setup, conducted experiments, and experimental results. Finally, Section 5 concludes the paper.

2. Related Work

The literature already reports several interesting crowd-sourcing based solutions, which are mostly based on offline or online surveys for infrastructure monitoring and feedback on different public services [11]. The majority of the recent solutions rely on smartphones and other hand-held devices by developing smart applications that allow users to give feedback on the infrastructure and services. For instance, Rapousis et al. [12] proposed QoWater, a client-to-server architecture-based mobile application allowing mobile users to give feedback on water quality. Similarly, Santani et al. [13] proposed CommuniSense, a mobile phone application for crowd-sourcing to monitor road conditions in Nairobi. However, several challenges are associated with such applications for crowd-sourcing [14]. One of the key limitations of such surveys is the narrow scope, which means they can cover a limited number of people as, generally, people are found reluctant to install and use such mobile applications. A prime example of people's reluctance to such mobile applications is observed during COVID-19 when people showed concerns over such applications in terms of privacy, difficulty in usage, and battery consumption [6]. Moreover, it takes a lot of time to complete a survey and also needs to involve several human and other resources, thus, the frequency of such surveys is generally very low, as it is costly and unfeasible to frequently conduct these surveys. These challenges could be resolved by extracting people's feedback on infrastructure and public services. The literature already provides some hints on the effectiveness of social media for real-time monitoring and instant feedback on different services. For instance, Wan and Paris [15] explored the potential of social media as a source of feedback

2024, Vol. 1, Article ID. 2024.0006
https://doi.org/10.69709/CAIC.2024.133109

on government services by analyzing citizens' opinions in a social media text.

Water quality analysis is one of the key applications that recently got the attention of the community. To this aim, several interesting frameworks have been introduced. The majority of the existing works aim at sentiment analysis of social media posts to extract people's opinions on water quality. For instance, Lambert [16] proposed a sentiment analysis framework for analyzing users' feedback and perception of tap water quality. Similarly, Li et al. [17] performed sentiment analysis on social media posts about recycled water in China. Jiang et al. [18], on the other hand, analyzed the public's opinion on large hydro projects by performing sentiment analysis on relevant social media posts. To this aim, three different hydro projects in China were considered, and mixed opinions were noticed for the projects.

More recently, water quality analysis from social media posts has also been introduced in MediaEval 2021 [19]. The task involved the retrieval of relevant multimedia content describing water quality in an Italian region. A couple of interesting solutions, incorporating different types of available information, are proposed in response to the task. For instance, Hanif et al. [20] fine-tuned existing pre-trained deep-learning models namely VGGNet and BERT for retrieving relevant visual and textual content, respectively. Overall better results are reported for textual content. Several limitations have been observed of the visual data provided in the dataset. For instance, images are available for very few posts. Moreover, the majority of the available images are not relevant. The issue is reported by Ayub et al. [21] and rather focused on textual content by employing three different NN models including BERT, RoBERTa, and a custom LSTM both individually and jointly in a naive late fusion scheme. This work employs a very basic scheme of fusion by treating all the models equally in the final decision, in contrast to merit-based weight assignment.

Despite the initial efforts in the domain, several interesting aspects of water quality analysis and automatic analysis of people's feedback on public services and infrastructure remain unexplored. For instance, the majority of the initial efforts are based on sentiment analysis without extracting meaningful information from the content itself. The domain also lacks a large-scale benchmark dataset. To this aim, a large-scale benchmark dataset has been collected and annotated on water quality analysis. The text classification framework is also extended with topic modeling to automatically extract key water-related issues discussed in social media.

3. Methodology

Figure 1 provides the block diagram of the proposed system. As can be seen, the proposed system is composed of five steps. In the first step, a large number of Tweets have been collected. In the next step, these tweets are annotated in a crowd-sourcing study. The annotated dataset is then used to train/fine-tune Large Language Models (LLMs) to classify tweets into relevant and non-relevant. In the fourth step, several merit-based fusion techniques are used to combine the classification scores obtained with the individual models. In the final step, topic modeling techniques are used to identify topics in the relevant tweets. In the next subsections, a detailed description of each step has been provided.

3.1. Data Collection, Cleaning, and Annotation

A crawler was developed for data collection, capable of continuously gathering data from different outlets of social media. As proof of concept, in the current implementation, data was collected from Twitter only. To this aim, a Python package namely Tweepy[1] was used with different relevant keywords, such as *waterpollution*, *water*, *watercrisis*, *watersmell*, *drinkingwater*, *watercolour*, *cleanwater*, *waterquality*, *plasticpollution*, *savewater*, *waterislife*, *cleantheocean*, *plasticocean*, *endplasticpollution*. The list of keywords is prepared in a data-driven manner by picking the keywords used in social media posts, blogs, newspapers etc., .Effort was made to include as many relevant keywords as possible to the list to collect relevant and quality tweets. This resulted in a large collection of tweets, which were saved in a CSV file. After data collection, all the collected tweets were manually annotated by involving multiple volunteers in a crowd-sourcing activity. Before the annotation, the collected data is manually checked to remove less informative tweets. For example, very short tweets were removed that did not have sufficient text or contained tags only. Duplicate entries in the file were also removed.

During the crowd-sourced activity, a total of 8,000 tweets were manually analyzed, which are annotated as relevant or non-relevant. To ensure the quality of the annotated data, each sample is checked by three different annotators and is labeled based on the majority votes. The participants of the crowd-sourcing activity were postgraduate students with sufficient knowledge of the domain.

[1] https://www.tweepy.org/

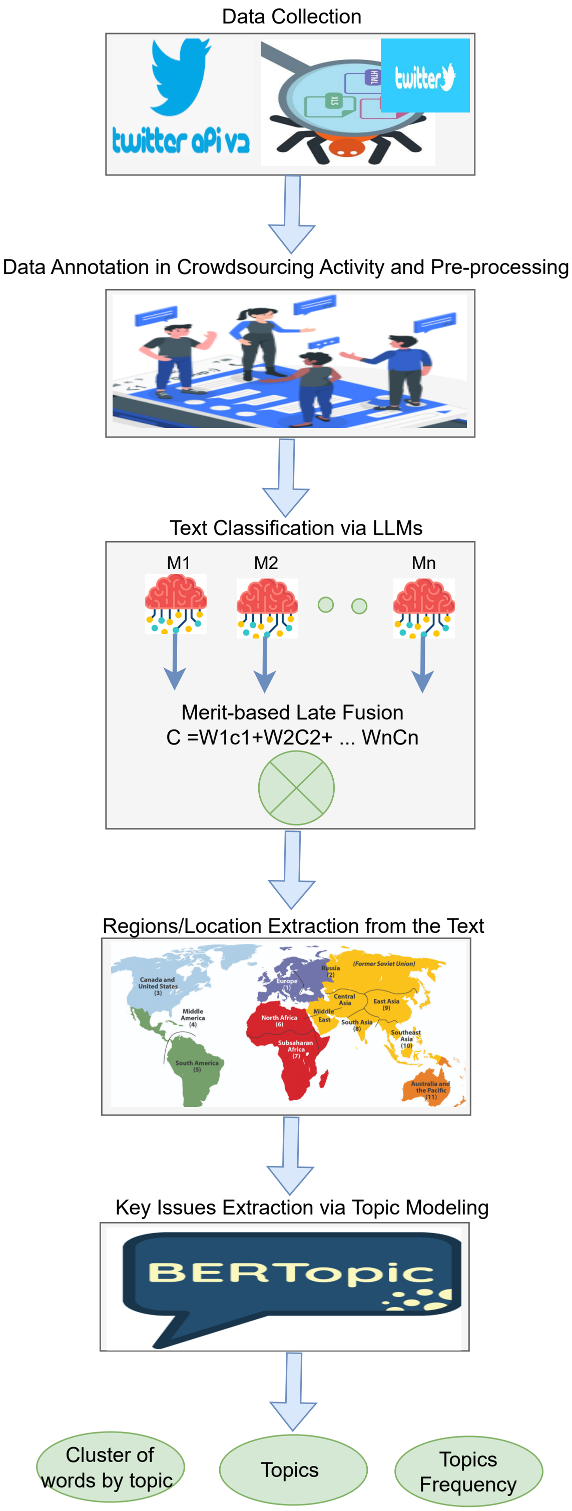

Figure 1: A block diagram of the proposed methodology.

3.2. Text Classification

For text classification, several LLMs were employed, both individually and jointly in a merit-based fusion technique to differentiate between relevant and non-relevant tweets. In the next subsections, a detailed description of the classification and fusion process is described.

3.2.1. Classification via Individual Models

In this work, the approach primarily relies on state-of-the-art transformer-based NLP models for the classification of tweets. In total, six different models are used. These models include the original BERT model, RoBERTa, ALBERT, DistilBERT, GPT, and Meta-LLAMA. The selection of these models is motivated by their proven performances in similar tasks, and we believe the evaluation of these models will provide a baseline for future work in the domain. A brief overview of these models is provided below.

- **BERT**: It is one of the state-of-the-art NLP algorithms that have been widely used for a diversified list of NLP applications. Its ability to read/learn in both directions makes it a preferred choice in different text-processing applications. Several implementations of BERT are available. In this work, Tensorflow implementation has been used. The model is composed of 12 layers and attention heads, and 110 million parameters. The loss function is based on the Binary Cross entropy loss function while the Adaptive Moments (Adam) optimizer is used in the experiments.
- **RoBERTa**: RoBERTa is another state-of-the-art transformer-based NLP model, and it uses self-attention for processing and generating contextualized representations of input text. One of the key advantages of RoBERTa over BERT is its training on a larger dataset and the use of a dynamic masking technique, allowing the model to learn robust and generalizable representations of words. In this work, the model was fine-tuned on our dataset using the Adam optimizer with a binary cross-entropy loss function.
- **ALBERT**: It is a modified version of BERT with fewer memory requirements. ALBERT has a reduced number of parameters mainly due to factorized embedding parameterization and cross-layer parameter sharing. In this first technique, the large vocabulary embedding matrix is decomposed into two small matrices, separating the size of the hidden layers from the size of the vocabulary embedding. The cross-layer parameter sharing, on the

other hand, prevents an increase in the number of parameters with the depth of the model.

- **DistilBERT**: DistilBERT is another variant of the BERT model aiming at applications with less computational and memory requirements. The concept of knowledge distillation is adopted during pre-training allowing a significant reduction in parameters without a significant impact on the performance of the model.
- **GPT**: Generative pre-trained transformer (GPT) models represent a family of Neural Network (NNs)-based language prediction models built on the Transformer architecture [22]. These models are pre-trained on a huge volume of diverse text data. Currently, GPT is available in different versions. However, the first version of the model was introduced in 2018 by Open AI [22]. In this work, GPT version 3.5 turbo was used, which is composed of 175 billion parameters, being significantly higher than the number of parameters used in its previous versions and other transformers, such as BERT. In this work, prompt engineering for the classification of tweets through GPT 3.5 was used.
- **Meta-LLAMA**: Large Language Model Meta AI (LLaMA) is also a family of pre-trained LLMs. Similar to GPT, multiple versions of LLAMA are available having 7B to 70B parameters. In this work, LLAMA 2 was used, which is an improved version of the base model LLAMA. Similar to the base model, LLAMA 2 is built on the Google transformer architecture with several interesting changes and improvements. For example, the RMSNorm pre-normalization, a SwiGLU activation function, and multi-query attention instead of multi-head attention and AdamW optimizer. The key differences between LLAMA 2 and the original LLAMA include a higher context length (i.e., 4096 compared to 2048 tokens) and grouped-query attention instead of multi-query attention. Similar to GPT 3.5, the prompt engineering method for text classification with LLAMA was used.

3.2.2. Fusion of the Models

The fusion methods are based on a late fusion scheme, where the scores/posterior probabilities of the individual models are accumulated for the final decision using Equation (1). In the equation, $S_{m1}, S_{m2}, S_{m3}, ...S_{mn}$ represent the scores/posterior probabilities obtained through the 1st, 2nd, , 3rd, and nth model, respectively while $W_1, W_2, W_3, ...W_n$ are the corresponding weights assigned to these models.

$$S_f = W_1 S_{m1} + W_2 S_{m2} + W_3 S_{m3} + ... + W_n S_{mn} \quad (1)$$

The weights are assigned to the models on the basis of their performances. To this aim, several weight optimization/selection methods, including PSO, Nelder Mead, BFGS, and Powell method, are employed. These methods seek a set of variable values (i.e., $W_1, W_2, W_3, ...W_n$ in our case) optimizing an objective function under a set of constraints. In this case, the fitness/objective function is based on accumulative classification error obtained on a validation set using Equation (2). In the equation, A_{acc} represents the accumulative accuracy computed on the validation set. In this work, the goal is to find a set of weights to be assigned to the models that minimize the classification error.

$$e = 1 - A_{acc} \quad (2)$$

It has been noted that the same fitness function is used by all the weight optimization methods employed in this work. These methods use different mechanisms and have their own pros and cons. A brief overview of each method is provided below.

- **PSO**: Particle Swarm Optimization (PSO), which is a heuristic approach, has been widely used in the literature for different tasks. For instance, in several works, PSO has been used for the optimization of hyper-parameters of ML algorithms, such as the number of layers, batch size, number of neurons, etc., in LSTMs and CNNs [23,24]. Similarly, it has been also used for the hyper-parameter optimization of Federated Learning (FL) algorithms [25]. The literature also reports the effectiveness of the optimization technique in late fusion where the algorithm is used to assign optimal weights to the classifiers [26,27]. The algorithm solves the optimization problem in three steps, iteratively; starting from a random set of candidate solutions, where each candidate solution is called a particle. At each iteration, each particle keeps track of its personal and global best solution in the swarm. The particles adjust two parameters namely (i) velocity and (ii) the position. The velocity of a particle is adjusted based on its own experience and the information shared by the other particles in the swarm. The position of particles is adjusted based on their current position, velocity, and distances between their current positions and personal and global best. This process continues until a global optimum is obtained. The key limitations of the method include a slow convergence rate, especially in high dimensional problems, and entrapment in local minima. Being one of the key optimization algorithms, PSO implementation is available in several libraries. In this work, the open-source library namely pyswarm[2] was used for the implementation of the algorithm.

- **Nelder Mead Method**: Similar to PSO, the Nelder Mead method has also been widely explored for different optimization tasks. For instance, Takenaga et al. [28] employed the method for computationally expensive optimization problems. Similarly, Ozaki et al. [29] used the algorithm for the hyperparameter selection/optimization of a CNN model. The method has also been widely used for the fusion of classification algorithms in different visual and NLP applications [9,30]. The method optimizes a set of variables leading to a minimum or maximum value of an objective function in a multidimensional space. To this aim, it uses a set of $n+1$ test points (solutions), which are arranged as a simplex. The method then estimates the behavior of the objective function at each test point for new test points, which replace the old ones in an iterative manner. In this work, a Python open-source library, namely, SciPy[3] was used for the implementation of the method.
- **Limited-memory Broyden Fletcher Goldfarb Shanno Algorithm (BFGS)**: Similar to PSO and Neldar Mead, BFGS and its variants have been proven very effective in different tasks, such as optimization hyperparameters of deep learning models and fusion. For instance, Saputro et al. [31] employed the algorithm for parameter estimation on a geographically weighted ordinal logistic regression model. Maria et al. [32] employed the method along with other optimization techniques for the fusion of inducers' scores for media interestingness prediction. BFGS, which is a local search optimization algorithm, belongs to the Quasi-Newton optimization family and aims at the optimization of the second-order derivative of the objective function. To obtain a set of optimal values, the algorithm computes the inverse of the Hessian matrix, used for multivariate functions. To this aim, the algorithm approximates the inverse, using a gradient that eliminates the need for inverse calculation at each step. One of the key limitations of the algorithm is its high memory requirement, which makes it impractical to compute the inverse of the Hessian matrix with a larger number of input parameters. To overcome this limitation, several variations of the algorithms have been proposed. For instance, Limited BFGS/LBFGS [33] is one of the variants of the algorithms with fewer memory requirements. In this work, despite not having a large number of inputs, the LBFGS implementation of the method was used.
- **Powell Method**: Powell method is another interesting optimization method that has been widely used for similar tasks. For instance, Maria et al. [32] and [9] employed the method for merit-based late fusion of classifiers for media interestingness and water quality analysis, respectively. Similar to PSO, several variations of the algorithm have been proposed in the literature. The algorithm seeks the local minima of the objective function. The objective function, which is a real-valid function with multiple inputs, doesn't need to be differentiable. The algorithm finds the minima in several steps starting with a random selection and evaluation of initial points/solutions, after which a list of parameters is randomly selected. A subset of the initial points with the minimum error is then selected as parents to produce offspring for the next generation. The children/new points are then evaluated in the fifth step and the process is repeated again from the third step until a global minima is found.

3.3. Regions Extraction

In this phase, different regions based on the locations associated with the tweets have been defined. This enables the analysis of the water quality or water-related issues in different regions of the world, as each region may have specific issues. It has been noted that this step is added to facilitate region-wise topic modeling, where the aim is to extract keywords used in water-related tweets from different parts of the world. To this aim, the location addresses associated with each tweet are fed into ChatGPT to identify the corresponding countries by mapping the addresses to the respective countries. To ensure the quality of the mapping, the identified countries and the associated addresses are meticulously verified. To further enhance the accuracy of the data, filtering techniques to specific locations were applied. For example, in cases where the user's location included the address 'Florida, FL,' it was replaced with 'USA'. This replacement was applied wherever the specified keyword was encountered. As a result, it became possible to successfully extract and verify 4707 accurate locations. The countries list was then provided to ChatGPT to expand the geographical scope by translating the unique countries into regions using ChatGPT.

[2] https://pyswarms.readthedocs.io/en/latest/
[3] https://scipy.org/

2024, Vol. 1, Article ID. 2024.0006
https://doi.org/10.69709/CAIC.2024.133109

3.4. Topic Modeling

The final component of the methodology is based on BERTopic [34], which is a state-of-the-art topic modeling technique. One of the key advantages of topic modeling is its ability to quickly discover the hidden topical patterns present in the data. These hidden patterns could result in meaningful insights leading to useful data-driven decisions. In this work, the aim was to automatically extract the hidden topical patterns in the water-related tweets, to identify the key water-related issues and concerns expressed over the water quality in the tweets.

The algorithm used in this work extracts topics from Tweets in three different steps, starting from converting the tweets into embeddings, then reducing the dimensionality and clustering, and finally converting them into topics. The embeddings are obtained by a pre-trained model namely Sentence-BERT. The dimensionality reduction and clustering are carried out through Uniform Manifold Approximation and Projection(UMAP) and HDBSCAN (Hierarchical DBSCAN), respectively. Finally, topics are extracted from the clustering using a modified form of TF-IDF (Term Frequency-Inverse Document Frequency) namely c-TF-IDF.

The algorithm brings several advantages. For instance, it clusters documents based on both lexical and semantic similarities. Moreover, BERTopic provides a library with several packages allowing more accurate and better visualization of the clusters, topics, and probabilities. It also comes with a few limitations. For instance, its assumption that each document/tweet contains only one topic is its main limitation, though it is possible to have Tweets with multiple topics. It is important to note that pre-processing was also performed in addition to data cleaning before topic modeling. For instance, short and stop words, numbers, and alphanumeric characters were removed, which allowed the removal of irrelevant frequently used words.

4. Experiments and Results

4.1. Dataset

The final dataset, after removing less informative tweets during the manual analysis and annotation, contains a total of 7,930 tweets. Among these, 5,728 tweets are annotated were irrelevant while the remaining 2,202 tweets were classified as relevant. The dataset has been divided into three subsets namely (i) training, (ii) test, and (iii) validation set using a ratio of 70%, 20%, and 10%, respectively. The validation set is used for the computation of the classification error for the fitness function of the fusion methods. Table 1 provides some sample relevant and irrelevant tweets from the dataset.

4.2. Experimental Results

The objectives of this work are multi-fold. On one side, the aim is to extract flood-related tweets, and on the other hand, the focus is to automatically extract keywords from the relevant tweets. It is highly possible that each country/region may have different water-related issues than others, thus, the interest is also in keywords/topics of tweets tweeted from a specific country/region. To achieve these objectives, the following experiments are performed.

- Evaluation of the performance of several state-of-the-art LLMs individually.
- Fusion of the classification scores obtained through the individual models in a merit-based fusion framework by employing several weight selection/ optimization methods.
- Topic modeling on all the relevant tweets. This will allow to highlight key global water-related issues.
- Topic modeling of the collection of tweets tweeted from a specific country/region. This will provide the opportunity to highlight the water-related issues specific to a particular region.

In the next subsections, a detailed analysis of the results of all the experiments has been provided.

4.2.1. Text Classification Results

Table 2 provides the results of the first experiments, where the performance of several LLMs in the application is evaluated. It is noted that for GPT and LLAMA-2, the prompt engineering method with a few-shot (5-shot and 10-shot) classification setting without fine-tuning the models is used. The highest F1-score of 0.768 is obtained with the original BERT model. As can be seen, overall similar results are obtained for BERT and its different variants and XLNET. This indicates that these models respond similarly to the data. However, the lowest results are observed with Meta-LLAMA-2 with 5-shots and 10-shots. One of the potential reasons for the lowest performance of the model is the few-shot learning, as the model may have limited generalization to classify the samples from the seen examples.

Table 3 reports the results of the fusion experiment, where the classification scores of the best-performing individual models are combined in a merit-based fusion scheme. In this experiment, two experimental settings were considered. In the first case, the classification scores obtained from the top 5 performing models were combined, including BERT, RoBERTa, DistilBERT, ALBERT,

2024, Vol. 1, Article ID. 2024.0006
https://doi.org/10.69709/CAIC.2024.133109

Table 1: Sample Tweets from the dataset.

Relevant Samples	Irrelevant Samples
We have been receiving water of the worst quality from past 6 months. I want to bring this situation to your notice and solve this problem ASAP. Water is a basic need. Area : Adarsh Nagar, Bahadurgarh	One of the most popular urban beaches in Gran Canaria, Las Canteras is a two-kilometer ribbon of sand caressed by warm and calm water.
Drinking contaminated water can transmit diseases and back in 2017 nearly 1.6 million people died from diarrheal diseases. 1/3 of those were children under the age of 5. #climatecrisis #water	Wondering about #books about #water sports (canoeing, sailing, yachting, scuba diving, etc.)? Check out call number range
The landmark research blames chemical #pollution from plastics, farm fertilisers and pharmaceuticals in the #water. Previously, it was thought the amount of #plankton had halved since the 1940s, but the #evidence gathered by the Scots suggests 90% has now vanished.	Hope people leave water out in their gardens or balcony in any containers for all the beautiful wildlife x #water #wildlife #thirsty #animals x
In face of recurring drought, cities seek security in wastewater recycling projects #security #projects #recycling #wastewater #water Removing pollution from water using water shaping tech #sketchup #depollution #watershaping #waterpollution	In a larger portion of cases, #carpet #damage is treated efficiently and all the defects are repaired. Professional services take care of all the #Water #Damage #Restoration Sunshine Coast.
The privatisation of water and power has been one of the biggest rip-offs of the British public in modern times. Time to jail those profiteering through pollution of our rivers and waterways! #water #corporategreed #utilities	The theme this time is ”Water from Japanese restaurants”. Is it true that there are many paid shops outside Japan? The popular article has exceeded 650pv. Is water free at Japanese restaurants?
Life without water is impossible. Save water. Save life. With every little drop, a day less to live on Earth.	Your body depends on #water to survive. Every cell, tissue, and organ in your body needs water to work properly and for overall good health. Learn how to ensure you stay hydrated, and why it is important to do so, here in family doctor
Drinking contaminated #water can be harmful to one’s health. #Cholera, #diarrhea, #dysentery, and #typhoid are just a few of the ailments it can induce.	We’ve worked with TheMixUK to explain what support is available for those struggling to pay for the increasing price of #Fuel and #Water bills. Take a read of the article here
Clean Water is a necessity to daily life. Empower economically disadvantaged small communities to develop and sustain clean water supplies.	Visit Central Florida Water Ski Sweepstakes

and XLNET while in the second experiment, the top 2 models namely BERT and ALBERT were considered.

Overall, there is a slight improvement in the results of the fusion compared to the best-performing individual model. Generally, fusion leads to an improvement in the F1 score. However, the limited improvement in this case could be due to the complexity of the dataset or the fewer variations in the individual models’ results. These fusion techniques improve results significantly when the individual models respond differently to the data. As far as the comparison of the fusion methods is concerned, no significant differences have been observed. However, the performance of all the methods is higher when the top 2 best-performing models are used in the fusion compared to the top 5 models. In the case of top models, though there is no significant difference, the slightly lower performance could be due to the low-performing models that could adversely affect the performance of the fusion methods.

2024, Vol. 1, Article ID. 2024.0006
https://doi.org/10.69709/CAIC.2024.133109

Table 2: Experimental results of the individual LLMs.

LLM	F1-score
BERT	0.7686
ALBERT	0.7636
DistilBERT	0.7491
ROBERTA	0.7541
XLNET	0.76
Gpt-3.5 (5-shot)	0.7146
Gpt-3.5 (10-shot)	0.7246
Meta-LLAMA2 (5-shot)	0.5832
Meta-LLAMA2 (10-shot)	0.5876

Table 3: Evaluation of the fusion methods.

Fusion Method	F1-score	
	Top 2 (BERT and ALBERT)	Top 5
Simple Averaging	0.770	0.7630
PSO	0.772	0.770
Nelder Mead Method	0.776	0.771
Powell Method	0.7713	0.7680
BFGS	0.77	0.7687

4.2.2. Location Extraction and Topic Modeling Analysis

In the topic modeling, two different experiments were conducted. In the first case, hidden topical patterns were analyzed and explored within the complete collection of water-related tweets in the test set. Figure 2 provides the top 10 topics and the corresponding words extracted from the collection of relevant tweets through BERTopic. As can be seen, most of the topics and associated words are very relevant. The issues highlighted by the algorithm from the tweet collection include sanitation & access to water, pollution, plastic pollution in water reservoirs, saving and utilization of rainwater, irrigation & drought issues, environmental factors, filtering drinking water, heatwaves, chemicals and tap water, etc.,

In the second experiment, the collected relevant tweets were divided into regions, which provided the opportunity to discover topics in tweets relevant to or tweeted from certain regions. This experiment helps to discover people's concerns about this important topic of water-related issues including both local and global issues. As a first step, the country names were extracted from the addresses associated with relevant tweets using ChatGPT. This resulted in a long list of countries from where water-related tweets were tweeted. It was observed that very few tweets were recorded from certain countries. For example, the collection of relevant tweets contains a single tweet from Slovenia, Mozambique, El Salvador, and Grenada. To ensure a sufficient number of tweets from each country, only those countries from which at least 70 tweets were tweeted were considered. It was noted that there is no scientific basis for this threshold (i.e., a minimum of 70 tweets per country). The focus was to ensure a sufficient number of countries on the list, while also guaranteeing enough tweets from each country for the analysis. Figure 3 provides the country-wise distribution of the relevant tweets in the dataset. A large portion of the tweets originated from the United States and the United Kingdom. This also indicates the interest of the people from these countries in this important societal challenge.

Figure 4 provides the list of topics extracted from the tweets originating from different countries. Topic 0 to Topic 6 show the group of topics extracted from the tweet collections for Australia, Canada, India, Pakistan, South Africa, the United States, and the United Kingdom, respectively. Some of the topics and the associated words are less relevant compared to the others. For example, most of the words associated with topic 0, which is extracted from tweets originating from Australia, are not very relevant to water-related issues. However, on the other side, Topics 2 to Topic 6 are very relevant and helpful in highlighting the issues. For instance, Topic 2 stresses the careful usage of water in general and rainwater in particular. Topic 3 and Topic 4 are about drinking water in one of the provinces of Pakistan and South Africa, respectively. Similarly, Topic 5 is based on heatwaves, wildlife, and water pollution. Finally, Topic 6 also includes relevant keywords, such as clean water and droughts.

Topic modeling was also performed on different geographic regions by combining tweets from all the countries in the region. These regions are formed on the basis of the geographic locations of the countries. To this aim, the list of countries was provided to ChatGPT resulting in five regions including Asia, Africa, America, Oceania, and Europe. Similar to country-wise topic modeling, the regions having at least 70 tweets were included. Figure 5 provides the distribution of relevant tweets from each region. As can be seen in the figure, overall, a higher number of tweets originated from America, Europe, and Asia.

Figure 6 provides the summary of the topics extracted from the tweets originating from different regions. Topic 0 to topic 4 represent topics extracted from tweets originating from Africa, America, Asia, Europe, and Oceania, respectively. The majority of the topics and the associated keywords are very relevant to water quality, except

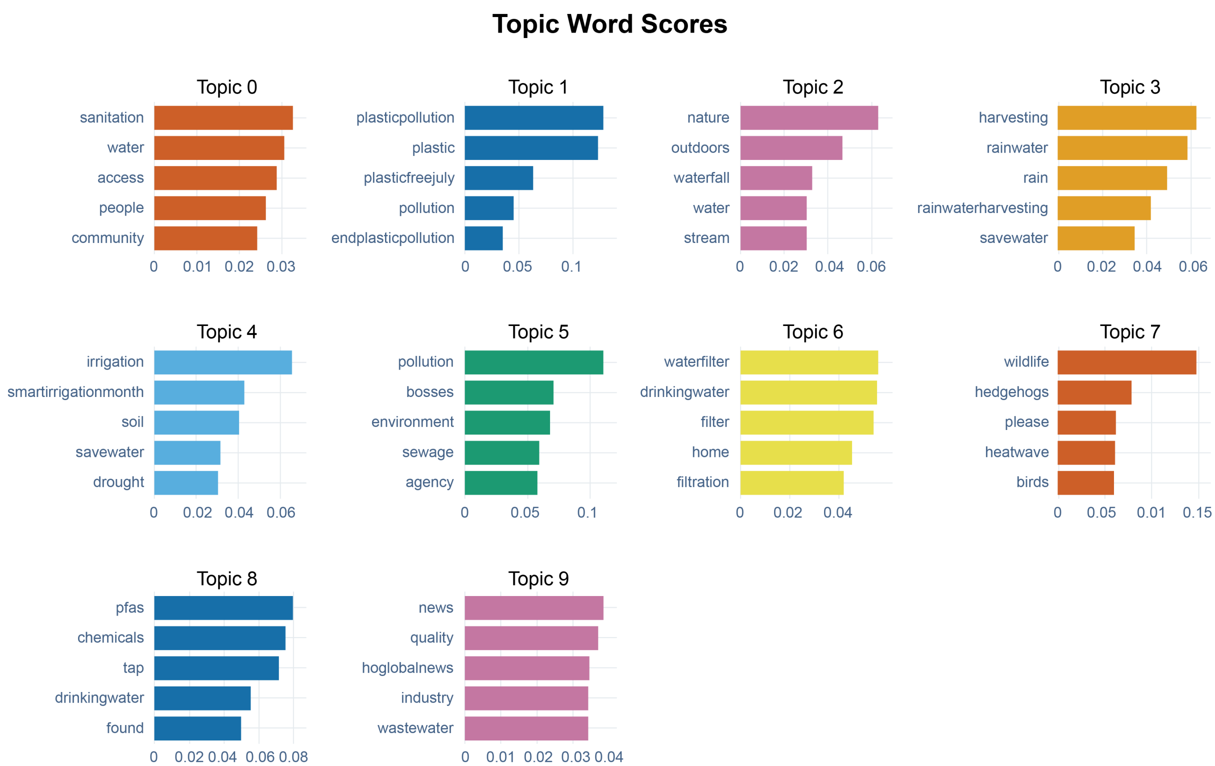

Figure 2: Top 10 hidden topic patterns extracted from the complete collection of relevant tweets.

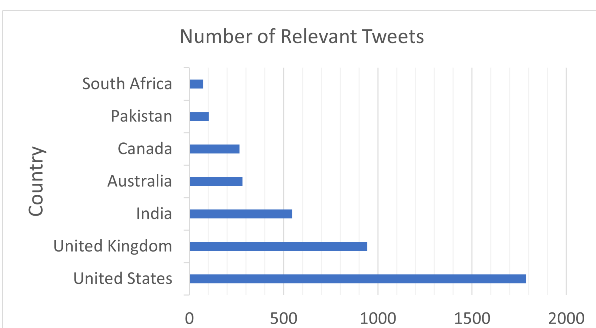

Figure 3: Country-wise distribution of the origin of the water-related tweets.

the topic extracted from the Oceania region. The topics are similar to what has been observed in the country-wise topic modeling, which indicates that the regions mostly have similar types of water-related issues or at least the topics/concerns are similar.

5. Conclusions

This paper presents a solution for automatic water quality analysis and the identification of hidden topic patterns in water quality-related tweets. The tweets were also analyzed which originated from different countries and regions, allowing global and regional water-related issues and concerns. In the text classification part, several LLMs were employed, both individually and jointly in a merit-based fusion framework. Overall, the models proved very effective in correctly classifying the tweets into relevant and irrelevant classes. No significant difference has been observed in the performance of the individual models. However, overall, better results are obtained when the models are jointly utilized. The topic modeling techniques also efficiently extracted relevant topics from the complete collection of tweets as well as tweets originating from specific regions and countries. The key topics discussed and extracted from tweets include 'pollution', 'plastic pollution', 'save water', drinking water, and 'sanitation', etc. Moreover, it was observed that a large portion of the relevant tweets are from certain countries, such as the United States and the United Kingdom.

In the future, the aim is to further extend the database by providing a deeper hierarchy of the labels (e.g., including the water pollution type and causes of the pollution). In the current form, only Twitter text is processed but the aim is to extend it to a multi-model framework by analyzing textual and visual data as visual information could be

2024, Vol. 1, Article ID. 2024.0006
https://doi.org/10.69709/CAIC.2024.133109

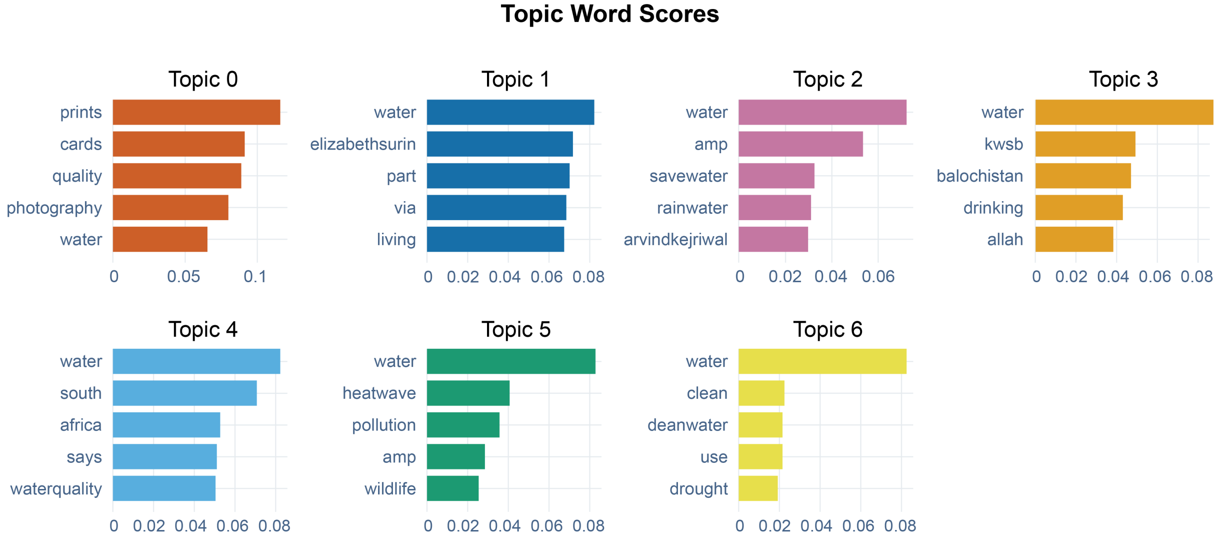

Figure 4: Country-wise topic modeling of the relevant tweets. Topic 0 to topic 6 represents topics extracted from tweets originating from Australia, Canada, India, Pakistan, South Africa, the United States, and the United Kingdom, respectively.

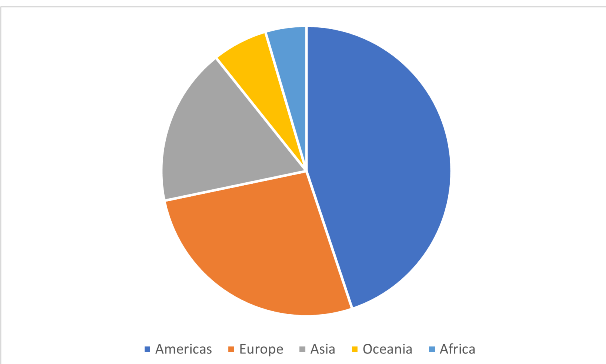

Figure 5: A region-wise distribution of the origin of the water-related tweets.

very effective in this application. Moreover, the aim is to extend the annotations to location information to prepare it for other NLP tasks, such as Named Entity Recognition.

List of Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
GPT	Generative Pre-trained Transformer
IoT	Internet of Things
LLM	Large Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
TF-IDF	Term Frequency–Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection

Author Contributions

Conceptualization, K.A. and M.A.A.; methodology, K.A. and M.T.Z.; software, I.K., H.N., M.T.Z.; validation, N.A., M.A.A.; formal analysis, K.A., N.A., M.A.A.; investigation, H.N., I.K., N.A.; resources, N.A. and K.A.; data curation, I.K., H.N. and M.T.Z.; writing—original draft preparation, K.A. and I.K.; writing review and editing, K.A. and N.A.; visualization, I.K. and M.A.A.; supervision, N.A, and K.A.; project administration, K.A. and N.A. All authors have reviewed and approved the publishe diversion of the manuscript.

Availability of Data and Materials

The data and other relevant materials will be made available upon request.

Conflicts of Interest

The authors declare no conflicts of interest regarding this manuscript.

Funding

No external funding was received for this project.

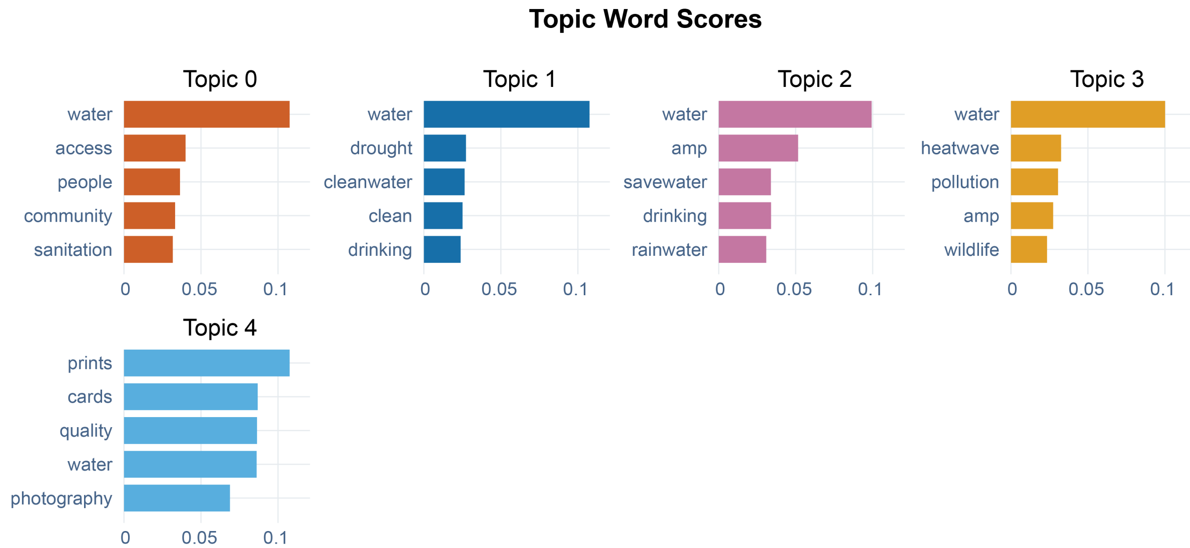

Figure 6: Region-wise topic modeling of the relevant tweets. Topic 0 to topic 4 represent topics extracted from tweets originating from Africa, America, Asia, Europe, and Oceania, respectively.

Acknowledgments

The authors acknowledge the efforts of the volunteers who helped in data annotation and other activities. The authors confirm that no content in this manuscript was generated using artificial intelligence (AI) tools and that they take full responsibility for the accuracy and integrity of the work.

References

[1] V. Marinakis, A. G. Papadopoulou, G. Anastasopoulos, H. Doukas, and J. Psarras, "Advanced ICT platform for real-time monitoring and infrastructure efficiency at the city level," in *Proc. 2015 6th Int. Conf. Inf. Intell. Syst. Appl. (IISA), IEEE*, Corfu, 2015, pp. 1–5. [CrossRef]

[2] V. Gagliardi et al., "Satellite remote sensing and non-destructive testing methods for transport infrastructure monitoring: Advances, challenges and perspectives," *Remote Sens.*, vol. 15, no. 2, 2023, Art. no. 418. [CrossRef]

[3] A. M. Helmrich et al., "Opportunities for crowdsourcing in urban flood monitoring," *Environ. Model. & Softw.*, vol. 143, 2021, Art. no. 105124. [CrossRef]

[4] M. S. D. P. Nayak and K. A. Narayan, "Strengths and weaknesses of online surveys," *J. Humanit. Soc. Sci.*, vol. 24, no. 5, pp. 31–38, 2019. [View Online]

[5] C. Andrade, "The limitations of online surveys," *Indian J. Psychol. Med.*, vol. 42, no. 6, pp. 575–576, 2020. [CrossRef] [PubMed]

[6] K. Ahmad et al., "Global user-level perception of covid-19 contact tracing applications: Data-driven approach using natural language processing," *JMIR Form. Res.*, vol. 6, no. 5, 2022, Art. no. e36238. [CrossRef] [PubMed]

[7] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and P. Halvorsen, "Social media and satellites: Disaster event detection, linking and summarization," *Multimed. Tools Appl.*, vol. 78, no. 3, pp. 2837–2875, 2019. [CrossRef]

[8] D. Westerman, P. R. Spence, and B. Van Der Heide, "Social media as information source: Recency of updates and credibility of information," *J. Comput. Mediat. Commun.*, vol. 19, no. 2, pp. 171–183, 2014. [CrossRef]

[9] K. Ahmad, M. A. Ayub, J. Khan, N. Ahmad, and A. Al-Fuqaha, "Social media as an instant source of feedback on water quality," *IEEE Trans. Technol. Soc.*, vol. 4, no. 4, pp. 325–333, 2022. [CrossRef]

[10] B. E. Lopez, N. R. Magliocca, and A. T. Crooks, "Challenges and opportunities of social media data for socio-environmental systems research," *Land*, vol. 8, no. 7, 2019, Art. no. 107. [CrossRef]

[11] A. H. Alavi and W. G. Buttlar, "An overview of smartphone technology for citizen-centered, real-time and scalable civil infrastructure monitoring," *Future Gener. Comput. Syst.*, vol. 93, pp. 651–672, 2019. [CrossRef]

[12] N. Rapousis, M. Katsarakis, and M. Papadopouli, "Qowater: A crowd-sourcing approach for assessing the water quality," in *Proc. 1st ACM Int. Workshop Cyber-Physical Syst. Smart Water Netw.*, Seattle, 2015, pp. 1–6. [CrossRef]

[13] D. Santani et al., "Communisense: Crowdsourcing road hazards in nairobi," in *Proc. 17th Int. Conf. Hum. Com-*

put. Interact. Mob. Devices Serv., Copenhagen, 2015, pp. 445–456. [CrossRef]

[14] P. Sillberg, J. Gronman, P. Rantanen, M. Saari, and M. Kuusisto, "Challenges in the interpretation of crowdsourced road condition data," in *Proc. 2018 Int. Conf. Intell. Syst. (IS), IEEE*, Funchal, 2018, pp. 215–221. [CrossRef]

[15] S. Wan and C. Paris, "Improving government services with social media feedback," in *Proc. 19th Int. Conf. Intell. User Interfaces*, Haifa, 2014, pp. 27–36. [CrossRef]

[16] L. H. Lambert and C. Bir, "Evaluating water quality using social media and federal agency data," *J. Water Health*, vol. 19, no. 6, pp. 959–974, 2021. [CrossRef] [PubMed]

[17] L. Li, X. Liu, and X. Zhang, "Public attention and sentiment of recycled water: Evidence from social media text mining in china," *J. Clean. Prod.*, vol. 303, 2021, Art. no. 126814. [CrossRef]

[18] H. Jiang, P. Lin, and M. Qiang, "Public-opinion sentiment analysis for large hydro projects," *J. Constr. Eng. Manag.*, vol. 142, no. 2, 2016, Art. no. 05015013. [CrossRef]

[19] S. Andreadis et al., "Watermm: Water quality in social multimedia task at mediaeval 2021," in *Proc. MediaEval 2021 Workshop*, Online, 2021. [View Online]

[20] M. Hanif, A. Khawar, M. A. Tahir, and M. Rafi, "Deep learning based framework for classification of water quality in social media data," in *Proc. MediaEval 2021 Workshop*, Online, 2021. [View Online]

[21] M. A. Ayub, K. Ahmd, K. Ahmad, N. Ahmad, and A. Al-Fuqaha, "NLP techniques for water quality analysis in social media content," in *Proc. MediaEval 2021 Workshop*, Online, 2021. [View Online]

[22] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI*, 2018. [View Online]

[23] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020. [CrossRef]

[24] T.-Y. Kim and S.-B. Cho, "Optimizing cnn-lstm neural networks with pso for anomalous query access control," *Neurocomputing*, vol. 456, pp. 666–677, 2021. [CrossRef]

[25] B. Qolomany, K. Ahmad, A. Al-Fuqaha, and J. Qadir, "Particle swarm optimized federated learning for industrial iot and smart city services," in *Proc. GLOBECOM 2020–2020 IEEE Glob. Commun. Conf., IEEE*, Taipei, Taiwan, 2020, pp. 1–6. [CrossRef]

[26] K. Ahmad, M. L. Mekhalfi, N. Conci, F. Melgani, and F. D. Natale, "Ensemble of deep models for event recognition, ACM Transactions on Multimedia Computing," *Commun. Appl. (TOMM)*, vol. 14, no. 2, pp. 1–20, 2018. [CrossRef]

[27] M. Qaraqe, M. Usman, K. Ahmad, A. Sohail, and A. Boyaci, "Automatic food recognition system for middle-eastern cuisines," *IET Image Process.*, vol. 14, no. 11, pp. 2469–2479, 2020. [CrossRef]

[28] S. Takenaga, Y. Ozaki, and M. Onishi, "Practical initialization of the nelder–mead method for computationally expensive optimization problems," *Optim. Lett.*, vol. 17, no. 2, pp. 283–297, 2023. [CrossRef]

[29] Y. Ozaki, M. Yano, and M. Onishi, "Effective hyperparameter optimization using nelder-mead method in deep learning," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, pp. 1–12, 2017. [CrossRef]

[30] M. T. Zamir, M. A. Ayub, J. Khan, M. J. Ikram, N. Ahmad, and K. Ahmad, "Document provenance and authentication through authorship classification," in *Proc. 2023 1st Int. Conf. Adv. Innov. Smart Cities (ICAISC), IEEE*, Jeddah, 2023, pp. 1–6. [CrossRef]

[31] D. R. S. Saputro and P. Widyaningsih, "Limited memory broyden-fletcher-goldfarb-shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR)," in *AIP Conference Proceedings*, Melville: AIP Publishing, 2017, vol. 1868. [CrossRef]

[32] M. Shoukat, K. Ahmad, N. Said, N. Ahmad, M. Hassanuzaman, and K. Ahmad, "A late fusion framework with multiple optimization methods for media interestingness," *arXiv*, 2022. [CrossRef]

[33] G. Yuan, Z. Wei, and Y. Wu, "Modified limited memory bfgs method with nonmonotone line search for unconstrained optimization," *J. Korean Math. Soc.*, vol. 47, no. 4, pp. 767–788, 2010. [CrossRef]

[34] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv*, 2022. [CrossRef]