Social Media as an Instant Source of Feedback on Water Quality

K. Ahmad, M. A. Ayub, K. Ahmad, J. Khan, Nasir Ahmad, and A. Al-Fuqaha, Senior Member, IEEE

Abstract—This article focuses on an important environmental challenge: the measurement of water quality, by analyzing the potential of social media to be harnessed as an immediate source of feedback. The goal of the work is to automatically analyze and retrieve social media posts relevant to water quality, with particular attention to posts describing different aspects of water quality, such as color, smell, taste, and water-related illnesses. To this aim, we propose a novel framework incorporating different preprocessing, data augmentation, and classification techniques. We use three neural networks (NN) architectures for our framework, namely: 1) bidirectional encoder representations from transformers (BERTs); 2) robustly optimized BERT pretraining approach (XLM-RoBERTa); and 3) a custom long short-term memory (LSTM) model. These are employed in a merit-based fusion scheme. For merit-based weight assignment to the models, several optimization and search techniques are compared, including a particle swarm optimization (PSO), genetic algorithm (GA), brute force (BF), Nelder–Mead, and Powell’s optimization methods. We also provide an evaluation of the individual models where the highest F1-score of 0.81 is obtained with the BERT model. Overall, in merit-based fusion, better results are obtained with BF achieving an F1-score of 0.852. We also provide a comparison against existing methods, where a significant improvement for our proposed solutions is obtained. We believe such a rigorous analysis is key to the field of water quality research.

Index Terms—Bidirectional encoder representations from transformer (BERT), genetic algorithms (GAs), late fusion, natural language processing (NLP), particle swarm optimization (PSO), RoBERTa, water crisis, water pollution, water quality.

I. INTRODUCTION

Over the last decade, social media outlets have proven to be an effective source of communication and information dissemination. The capability of social media to engage large audiences worldwide makes them a preferred platform to discuss and convey concerns over different domestic and global challenges [1], [2]. The literature already reports their effectiveness in a diversified set of societal, environmental, and technological topics, such as food security [3], discrimination and racism [4], hate speech and crime [5], [6], public health [7], natural disasters [8], and technological conspiracies [9].

It is noticed that in their posts, social media users generally explicitly identify the regions having water-related issues along with the relevant information. Therefore, public authorities can use this information as valuable feedback on water distribution networks. However, generally, several challenges are associated with the extraction of relevant information from such informal sources. For instance, it is possible that the posts containing relevant or similar keywords do not represent the actual debates on water quality. In addition, manually filtering and analyzing large collections of social media posts are a tedious and time-consuming process.

Recently, machine learning (ML) and natural language processing (NLP) techniques have shown outstanding capabilities in similar applications. We believe ML and NLP techniques could also be employed in this important application to automatically analyze and filter water-related social media posts. The automatic acquisition/retrieval and analysis of such water quality analysis techniques will result in a significant reduction in time and effort spent on the process. To explore the potential of ML and NLP techniques in the water quality domain, in this work, we propose a classification framework incorporating several preprocessing, data augmentation, classification, and fusion techniques. The preprocessing techniques allow cleaning of the data by removing URLs, punctuation, etc. The data augmentation, where we used a back translation scheme, serves two purposes. It increases our training set, and also helped to balance the dataset by increasing the number of training samples in minority classes. For fusion, we employed both the naive fusion method by treating all the models equally as well as merit-based fusion schemes. In the merit-based fusion, five different weights selection/optimization techniques are used to assign weights to three different state-of-the-art architectures; namely, bidirectional encoder representations from transformer (BERT), XML-RoBERTa, and long short-term
The main contributions of the work can be summarized as follows.

1) We explore a relatively new application; namely, water quality analysis in social media posts, by proposing a complete framework starting from preprocessing, data augmentation, classification, and fusion techniques.

2) We evaluate multiple state-of-the-art models, both individually and jointly using several late fusion techniques.

3) We also demonstrate how the performance of the classification framework improves by considering the performance of individual models in a late fusion scheme using five different weight selection and optimization techniques.

The remainder of this article is organized as follows. Section II provides an overview of the related work. Section III provides a detailed description of the proposed methodology and fusion techniques employed in this work. Section IV provides the description of the dataset, evaluation metrics, and experimental results. Finally, Section V concludes the work.

II. RELATED WORK

Being an important factor in a healthy life, water quality monitoring and analysis have always been a concern for public authorities. To this aim, different strategies and sources of information, such as satellite imagery and crowd-sourcing platforms [10], are utilized. For instance, Galvin et al. [11] proposed a mobile application; namely, cyanobacteria assessment network (CyAN), to monitor, detect, and disseminate information about water quality in lakes using remotely sensed images. Similarly, Mohsen et al. [12] employed remote sensing techniques to monitor and analyze the quality of the water of Lake Burullus in Egypt in satellite imagery. Satellite imagery has also been employed in several other interesting works for water quality analysis in lakes in different parts of the world [13], [14]. However, satellite imagery provides a bird’s eye view and is generally used for large water reservoirs, such as lakes and dams.

On the other hand, crowdsourcing techniques help in obtaining more detailed, contextual, and localized information. The literature already reports several interesting crowdsourcing solutions for monitoring water quality. For instance, Rapousis et al. [10] proposed QoWater, a client-to-server architecture-based mobile application allowing mobile users to give feedback on water quality. Similarly, Jakositz et al. [15] proposed and conducted a competition-based crowdsourcing study for tap water quality monitoring. One of the main drawbacks of such platforms for crowdsourcing is the limited number of users.

This limitation could be addressed with social media. Social media outlets, such as Facebook, Twitter, and Instagram, provide access to a large number of users. Such platforms could be utilized to provide instant feedback on water quality [16], [17]. However, extracting meaningful information from such informal sources is very challenging.

Thanks to recent advancements in ML and NLP techniques, social media information could be automatically analyzed and filtered to extract relevant information. There are already some efforts in this direction. For instance, Lambert and Bir [18] performed sentiment analysis on users’ feedback in social media posts to obtain their opinions and perceptions of tap water quality. Li et al. [19], on the other hand, provide sentiment analysis of public opinions on social media on recycled water in China. Similarly, sentiment analysis of social media posts is also carried out by Do [20] where some basic NLP techniques, such as bag of words (BoW) and Naive Bayes, Bernoulli Naive Bayes, and logistical regression classifiers, are used.

More recently, water quality analysis in social media posts has also been introduced in a benchmark competition; namely, MediaEval [21]. In the task, participants were asked to develop automatic tools that differentiate between relevant (i.e., water quality discussions) and irrelevant Twitter posts. The task mainly focuses on tweet text; however, additional information in the form of images associated with the tweets and metadata were also included. In total, two teams, including our team, completed the tasks by introducing several interesting solutions. For instance, Hanif et al. [22] proposed a multimodal solution incorporating both images and text. For visual content, a pretrained model, VGGNet, is fine-tuned, while for textual features a BERT model is fine-tuned on the dataset. The authors also submitted the results of the visual information-based solution only. However, the results indicate lower performances for both solutions. On the other hand, considering the quality of the textual and visual content, our team [23] decided to focus on textual information only. To this aim, three different neural networks (NNs) models; namely, BERT, RoBERTa, and a custom LSTM, are employed both individually and jointly in a naive fusion scheme by treating all the models equally.

However, we believe merit-based fusion schemes could better exploit the potential of the models by assigning weights to the models based on their performance.

III. PROPOSED METHODOLOGY

Fig. 1 provides the block diagram of the proposed methodology. The proposed method can be roughly divided into three phases starting with a preprocessing and data augmentation phase where several strategies are used to clean and augment the data. After preprocessing, multiple NNs models are trained on the data. Finally, the classification scores obtained with the individual models are combined using several merit-based late-fusion schemes. In the next sections, we provide a detailed description of each phase.

A. Preprocessing and Data Augmentation

In the preprocessing and data augmentation phase, we employed different strategies to clean and increase the number of training samples. As a first step, we cleaned the text
by removing URLs, account handles, emojis, and unnecessary punctuation. After cleaning the data, we performed data augmentation using a text translation technique; namely, back-translation. The text translation approach naturally suits our application as the dataset is composed of both Italian and English tweets where Italian tweets are translated into English and added to the original training set. This technique helps in generating more training samples without disturbing the context of the data.

Data augmentation serves two purposes. First, it increased our training set. Second, it helped in balancing the training set by increasing the minority class. Moreover, along with data augmentation, we also used an upsampling technique to balance the dataset. Before the data augmentation process, our training set contained 5084 and 1073 samples in the negative and positive classes, respectively. During the data augmentation process, the positive class was increased to 3219 samples.

B. Model Training and Classification

In this work, we used three different state-of-the-art NN architectures that include a couple of transformers, namely, BERT [24], XLM-RoBERTa [25], and an LSTM. The choice of the models is motivated by their proven superior performance in text classification tasks. In the case of BERT and XLM-RoBERTa, we fine-tuned existing pretrained models while a customized model is trained for LSTM. Moreover, we note that in the current implementation, the hyperparameter values are selected by the hit and trial method. In the future, we will use some search methods to select the best combination of the hyperparameter values. The details of the models are provided as follows.

1) BERT-Based Solution: BERT is a multilayer encoder. In contrast to conventional NLP models, BERT relies on a bidirectional training mechanism by taking into account both the previous and next tokens. Such training capabilities allow it to better extract contextual information. In this work, we rely on a pretrained BERT model, which is fine-tuned on the water quality analysis dataset. The model is composed of 12 layers, 12 attention heads, and 110 million parameters. We note that necessary preprocessing of the cleaned training data is carried out, using Tensorflow libraries, to bring the data in the required form to be fed into the model. Since our dataset is composed of two classes only (i.e., binary classification task), we used the binary cross-entropy loss function along with the adaptive moments (Adam) optimizer. Table I summarizes the parameters setting of the BERT model used in this work.

2) XLM-RoBERTa-Based Solution: XLM-RoBERTa is a multilingual version of RoBERTa. RoBERTa itself is a modified version of BERT. The model is trained on a large-scale dataset covering text from 100 different languages using mask language modeling (MLM) objective. The multilingual nature of the model makes it a better choice for our application. As per the requirements of the model, the input text is tokenized before feeding into the model. The model is then fine-tuned on the water quality analysis dataset using an Adam optimizer with a binary cross-entropy loss function. A summary of the parameter settings of the model is provided in Table II.

3) LSTM-Based Solution: Our third model is based on the LSTM architecture. LSTM is a recurrent NN (RNN) with better memorizing pattern capabilities, which makes it a better choice for text classification compared to classical ML algorithms, such as decision trees, random forests (RFs), and support vector machines (SVMs). In this work, we used a customized model composed of three layers, including an input, LSTM, and output layer. Our model is composed of 491 713 trainable parameters. A summary of the parameter settings of the model is provided in Table III.
TABLE III
PARAMETER SETTINGS OF THE LSTM MODEL.

| Attribute       | Value |
|-----------------|-------|
| Total Layers    | 4     |
| Embedding Vector Space | 32    |
| LSTM Vector space | 64    |
| Cost Function   | Cross entropy |
| Number of Classes | 20   |
| Training Solver | Adam  |
| Mini-Batch Size | 64    |
| Activation      | Sigmoid |
| Drop out        | 0.1   |
| Learning Rate   | 0.001 |
| Epochs          | 20    |
| Validation data | yes   |

C. Fusion

For the fusion, we mainly rely on late fusion schemes where both naive and merit-based fusion methods are employed. Our baseline method is based on a simple aggregation of the classification scores obtained with all the models. In merit-based fusion, we deploy different optimization and search techniques to optimize the weights assigned to the models in fusion. We note that in the current implementation, we use a linear combination of the models in the late fusion using

\[ F_c = W_1C_1 + W_2C_2 + \cdots + W_mC_n \]  

(1)

Here, \( F_c \) represents the combined classification score obtained as an outcome of fusion while \( w_n \) denotes the weight assigned to the \( n \)th model whose score is represented by \( C_n \). In our case, \( n = 3 \). In the case of naive baseline fusion, all the models are assigned equal weights (i.e., \( W_1 = W_2 = W_n = 1/N \)). In the merit-based fusion, on the other hand, values of the weights are selected based on the optimization/search methods used in this work.

Moreover, in the weight optimization-based fusion methods, our fitness/objective function is based on the accumulative error computed on the validation set using

\[ E_{acc} = 1 - A_{acc}. \]  

(2)

Here \( A_{acc} \) represents the cumulative accuracy computed on the validation set using (3). In the equation, \( p_n \) represents the probability/score obtained with the \( n \)th model while \( x_n \) is the weight to be assigned to the \( n \)th model

\[ A_{acc} = x_1 * p_1 + x_2 * p_2 + \cdots + x_n * p_n. \]  

(3)

We note that the choice of the evolutionary methods for weight selection is based on their proven performance in weight optimization for the fusion. The details of the fusion methods used in this work are provided in the next sections.

1) Particle Swarm Optimization-Based Fusion: The use of the particle swarm optimization (PSO)-based method is motivated by its promising performance in similar tasks [26], [27]. The key concept of PSO is inspired by the social behavior of birds flocking and fish schooling, where the idea is to get benefit from the experience of each other in finding the best solution. To this aim, it starts with an arbitrary population of the possible solutions, which are termed as particles, and tries to iteratively optimize the potential solutions to satisfy a given constraint provided in the objective function. To find the best global minimum, the algorithm keeps track of the current and best position and velocity of each particle at each iteration, which is then updated in successive iterations.

In this work, each combination of weights is considered a potential solution whereas our objective function is based on the accumulative error \( (E_{acc}) \), computed by (2). It is important to mention that PSO is a heuristic solution, and the solution is not guaranteed to be optimal. However, the literature indicates that the solutions found by PSO are generally close to the optimal one. Table IV provides a summary of the pros and cons of PSO.

2) Genetic Algorithms-Based Fusion: Genetic algorithm (GA), which is inspired by Charles Darwin’s theory of natural evolution, has also been widely explored in the literature for similar tasks involving weight selection and optimization [26]. The basic idea behind GA is to incorporate the natural phenomena in search/selection problems by selecting the best one among the potential solutions at each iteration.

The GA-based search/selection process is composed of several phases. The process starts with a random population of individuals/potential solutions (i.e., randomly selected weight combinations). The algorithm then searches for the fittest individuals (i.e., weight combination) by evaluating the individuals/potential solutions against fitness criteria provided in the fitness function, iteratively. The process continues by employing crossover and mutation operations until the population convergences (i.e., no further improvement is possible). Crossover and mutation are key operations directly contributing to the performance of the algorithm. The former aims to push the population toward local minimum/maximum and the latter aims to explore the best among the candidate local minimum/maximum solutions.

In this work, similar to PSO-based fusion, our fitness criterion is based on the accumulative error \( (E_{acc}) \), which is computed on a separate validation set using (2).

3) Brute Force-Based Fusion: The brute force (BF) search, which is also known as an exhaustive search, tries all the possible solutions to find a satisfactory one for an underlying problem. The BF method brings several advantages. For instance, it guarantees the best solution by considering all the possible solutions before choosing the best one. Moreover, a simple working mechanism makes it a preferred solution for small problems in a diversified list of applications. However, on the other hand, there are several limitations of the method. For instance, the complexity of the method is very high in high-dimensional applications. Moreover, it takes a lot of time to find all the possible solutions for a high-dimensional application.

In this work, we used the method to find the best combination of the weights assigned to the models that minimize the classification error. This method suits our task as we have only three models and the method needs to find very few combinations for the selection of the best combination. In this work, we used an open-source Python library, namely, SciPy, for the implementation of the algorithm, which aims to find the grid...
point having the lowest value of the objective function [i.e., cumulative error defined by (2)].

4) **Powell’s Method-Based Fusion:** In this method, we rely on the evolutionary Powell’s method for the optimization of the weights to be assigned to the individual models during fusion. The method is inspired by the original method [28], however, a stochastic element is introduced. The method aims at the global minima of the objective function, which is based on a cumulative error in our case.

The algorithm works in several steps. The process starts by randomly selecting and evaluating a few points/solutions. In the second step, a list of parameters is selected in random order. In the third step, a portion of the previously evaluated points/solutions is used as parents by ensuring the selection of points with the lowest error. The algorithm then looks for the children in the next generation. Finally, all the children are evaluated in the fifth step and the process is repeated from the third step. It is important to mention that while searching for the minimum, the algorithm moves in one direction only until it finds the local minima. The algorithm moves in the other direction once the minimum is found in the current direction. Similar to the other methods used in this work, our objective/cost function is based on the cumulative error defined by (2). For the implementation of the method, we used a Python open-source library, namely, SciPy.

### IV. Results and Analysis

#### A. Dataset

For the evaluation of proposed solutions, we used a large-scale dataset introduced in a benchmark competition task, namely, “WaterMM: Water Quality in Social Multimedia” MediaEval 2021 [21]. The dataset is composed of a large collection of Twitter tweets tweeted in English and Italian from May 2020 to April 2021. The data are collected using English and Italian keywords related to water quality, color, pollution, and water-related illnesses. The main challenge lies in differentiating between water quality-related tweets and irrelevant tweets containing terms, such as water, floods, etc. Table V provides some sample tweets from both classes.

The dataset is manually annotated by analyzing the tweets under the guidelines provided by the authorities of the Eastern Alps River Basin District, who are responsible for hydrogeological defense in north–east Italy. Each tweet is annotated as either relevant or irrelevant to water quality. In total, the dataset is comprised of 10 000 tweets. The training and test sets are already separated by the task organizers. The training set is comprised of 8000 tweets and the test set covers 2000 tweets. The training set is imbalanced containing only around 17.18% of relevant tweets, which poses challenges in training AI models.

The dataset is manually annotated by analyzing the tweets under the guidelines provided by the authorities of the Eastern Alps River Basin District, who are responsible for hydrogeological defense in north–east Italy. Each tweet is annotated as either relevant or irrelevant to water quality. In total, the dataset is comprised of 10 000 tweets. The training and test sets are already separated by the task organizers. The training set is comprised of 8000 tweets and the test set covers 2000 tweets. The training set is imbalanced containing only around 17.18% of relevant tweets, which poses challenges in training AI models.

#### B. Evaluation Metrics

The evaluation of the proposed solutions is carried out in terms of three different metrics: namely, precision, recall, and F1-score. The performance of the models is assessed by comparing their predictions against the ground truth annotations. The models that achieve the highest F1-score are considered to be the most effective in differentiating relevant from irrelevant tweets.

---

**Table IV**

| **Method**      | **Pros**                                                                 | **Cons**                                                                                           |
|-----------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| PSO             | - Easy to understand and implement                                       | - May easily fall and stuck in local minima                                                     |
|                 | - Efficient and insensitive to scaling                                    | - Low convergence rate                                                                            |
|                 | - Fewer parameters                                                       | - It is a heuristic solution and does not guarantee a globally optimal solution                  |
|                 | - Suitable for concurrent parameters                                     |                                                    |
| GA              | - Robust to local minima                                                | - Computationally expensive                                                                      |
|                 | - Can be easily parallelized for concurrent processes                     |                                                    |
| BF              | - Provides a guaranteed best solution                                    | - Its complexity increases with an increase in the dimensionality of the problem                  |
|                 | - Applicable to several problems from different domains                  | - Very slow                                                                                        |
|                 | - Simple to understand and implement                                     |                                                    |
|                 | - Better suited for small problems                                       |                                                    |
| Powell’s Method | - The objective function does not need differentiable                   | - May not find local minima in many iterations                                                   |
| Nelder–Mead    | - Works with function evaluations only                                   | - Not efficient                                                                                    |
|                 |                                                                        | - May take a large number of iterations without many changes in function value                  |

---

n = 3, and the objective function is based on the cumulative error defined by (2). For the implementation of the method, we used a Python open-source library, namely, SciPy.
### TABLE V

| Sample Tweets from the Dataset |
|-------------------------------|
| **Relevant Samples**          |
| "Alta concentrazione di cloro, acqua non potabile nella zona di San Pietro Laneto." (Translation: High concentration of chlorine, non-drinking water in the San Pietro Laneto area) |
| "PPP P Shane please supply water #INeedWaterCrisis" |
| "I'm out of every 3 people on our planet do not have access to clean water." |
| "Se l'occidentale avesse problemi di acqua potabile, non ci metteremmo due secondi a depurare l'acqua del mare, lo stesso bisognerebbe fare con i panni del sud del mondo." (Translation: If the West had problems with drinking water, it would not take us seconds to purify the sea water, the same should be done with the countries of the southern hemisphere.) |
| "Almeno il 15% delle acque coltivate del pianeta subisce una carenza idrica non dovuta a vincoli idrogeologici ma causata da una C" (Translation: At least 15% of the planet’s cultivated land suffers from a water shortage not due to hydrological constraints but caused by a...) |
| "Never saw water supply project faces more delays." |
| "Grazie al monitoraggio della gravità sulla Terra viene possibile stimare la situazione dell’acqua dolce a livello planetario e per." (Translation: Thanks to the monitoring of gravity on Earth, it is possible to estimate the situation of fresh water at the planetary and per level.) |
| "RT @ContrattiFrame: Che cosa finisce nelle nostre tazze? E con quali impatti su #Salute e #Ambiente? Utilizzati come discariche dove smaltire." (Translation: RT @ContrattiFrame: What ends up in our #water? And with what impacts on #Health and #Environment? Used as landfill where to discard.) |
| "Acqua di nuovo potabile a Cabibbio e Muggio https://t.co/IVXhJi6H (Translation: Drinking water again in Cabibbio and Muggio https://t.co/IVXhJi6H)" |
| "Serra, ordinanza del commissario: non potabile la acqua proveniente dai serbatoi Scorratainat. e Timpane Tondo. #Calabria #calabrianatizi #acqua." (Translation: Serra, order of the commissioner: the water from the a ’Scorratainat. and Timpe Tondo’ tanks is not drinkable #Calabria #calabrianatizi #water.) |

| **Irrelevant Samples**          |
| "Neanché le 7 cod ho giA. cambiato una gennaia ad un collega sul raccordo sotto l’acqua." (Translation: Not even 7 and I have already changed a tire to a colleague on the fitting under the water.) |
| "Some one just threw a water bottle towards PPD" |
| "So3IPodcast Need one for my water bottle! Bottle your tears for me to water my plants best by 9AM Monday." |
| "So rivestiro una bottiglia d’acqua, mi sono fatta il bagno e allagato la cucina, t’bn." (Translation: I spilled a bottle of water, took a bath and flooded the kitchen, t’bn.) |
| "Un’acqua stasera ho un vegetariano e una che A. dieta e non mangia carboidrati e legumi QUINDI mangiamo una bottiglia di acqua naturale." (Translation: for dinner tonight I have a vegetarian and one who is on a diet and does not eat carbohydrates and legumes SO we will eat a bottle of still water.) |
| "Fresh water bottle." |
| "La cina stasera ho un vegetariano e una che A. dieta e non mangia carboidrati e legumi QUINDI mangiamo una bottiglia di acqua naturale." |
| "@ApolloVentuno Now? wonderful! Here we are still underground!" |

### TABLE VI

| Evaluation of the Individual Models in Terms of Microprecision, Recall, and F1-Score |
|---------------------------------------|
| Methods | Precision | Recall | F1-Score |
|---------|-----------|--------|----------|
| BERT    | 0.833     | 0.790  | 0.811    |
| XLM-RoBERTa | 0.81     | 0.579  | 0.687    |
| LSTM    | 0.886     | 0.640  | 0.743    |

### TABLE VII

| Evaluation of the Fusion Methods in Terms of Microprecision, Recall, and F1-Score |
|---------------------------------------|
| Methods | Precision | Recall | F1-Score |
|---------|-----------|--------|----------|
| Baseline (Equal Weights) | 0.873 | 0.760 | 0.813 |
| PSO     | 0.781     | 0.916  | 0.843    |
| GA      | 0.791     | 0.895  | 0.840    |
| Brute Force | 0.810     | 0.900  | 0.852    |
| Powell’s Method | 0.810 | 0.897 | 0.851 |
| Nelder-Mead Method | 0.862 | 0.807 | 0.834 |

C. Experimental Results

In this section, we provide a detailed description and analysis of the experimental results. We also provide comparisons against the methods proposed in the benchmark competition.

1) Evaluation of Individual Models: Table VI provides the experimental results of the individual models in terms of precision, recall, and F1 score. As can be seen, overall better results are obtained with the BERT model while surprisingly the least score is obtained with RoBERTa. However, there is no single winner in terms of all of the three metrics. For instance, the highest F1 score is obtained with the BERT model while LSTM achieved the highest score in terms of recall. The lower F1-Score for LSTM compared to BERT indicates a higher number of false-positive samples. This variation in these metrics is an indication of the variations in the performances of these models in both classes, which provides the basis for our fusion-based experiments.

2) Evaluation of the Fusion Methods: Table VII provides the experimental results of the fusion experiment. In this experiment, we evaluated several fusion methods, including a naive method by treating all the models equally and a merit-based fusion strategy where five different weight selection/optimization methods are employed to assign weights to the models.

As can be seen, interestingly, no significant improvement in the performance of the naive baseline fusion scheme over the best-performing individual model has been observed.
TABLE VIII

| Methods                  | Experimental Results |
|--------------------------|-----------------------|
|                          | F1-Score  | Methods  | F1-Score  |
| Hanif et al. [22]        | 0.371     | Astif et al. [23] | 0.794     |
| This Work (PSO)          | 0.843     | This Work (GA)   | 0.840     |
| This Work (BF)           | 0.852     | This Work (Powell’s Method) | 0.851     |
| This Work (Nelder–Mead Method) | 0.834    | -         | -         |

One of the possible reasons could be the low-performing model (i.e., XML-RoBERTa) as it simply aggregates the classification scores of all the models. These results provide the basis for our merit-based fusion where weights are assigned to the models based on their performance in the first experiment.

In merit-based fusion, significant improvement in terms of F1-score is obtained with all the methods over the best performing individual model and the naive baseline fusion method. This emphasizes the fact that merit-based weights should be assigned to contributing models in the fusion.

As far as the performance of the individual weights selection strategies for the merit-based fusion is concerned, better results are obtained with the BF-based method having an improvement of 0.001 and 0.009 over the Powell and PSO-based fusion methods, respectively. One of the potential causes of better results of BF is its ability to provide a guaranteed best solution, where it first searches for all possible combinations and picks the best weight combination that maximizes the performance. The literature suggests that the complexity of BF increases with an increase in dimensionality and consumes more computation power. However, in this work, we are considering just three models (i.e., dimension = 3) and thus, the method is better suited for this application in the current implementation. It is important to note that the difference in the performances of BF and other competing methods, such as PSO, GA, and Powell’s method, is negligible.

3) Comparisons Against the Existing Methods: We also provide comparisons against existing works proposed for the task in a benchmark competition; namely, MediaEval 2021 [21]. In total, two teams managed to complete the task. Our team also participated in the competition and obtained the highest scores. As can be seen in Table VIII, all the merit-based fusion techniques employed in this work obtained significant improvements over the existing solutions. Our best performing merit-based method, namely, BF-based fusion, obtained an improvement of 48.1% over the method proposed by Hanif et al. [22]. On the other hand, it obtained an improvement of 4.1% over the method proposed by our team [23] for the task in the competition, where we proposed a naive fusion method by treating all the models equally.

The significant improvement in the performance of the water quality analysis framework indicates the significance of merit-based fusion.

D. Lessons Learned

The lessons learned during this work can be summarized as follows.

1) Recently, water quality analysis got the attention of the research community, and several interesting solutions incorporating different sources of information have been proposed.

2) Crowdsourcing has been one of the potential solutions to obtain relevant feedback on water quality; however, it is a tedious and time-consuming process to obtain a sufficient number of participants. Social media outlets provide a better platform to involve a large number of volunteers in crowdsourcing for water quality analysis.

3) As demonstrated in this work, ML and NLP techniques allow automatic analysis and extraction of relevant information from large collections of social media posts.

4) The classification results are significantly improved by jointly employing the state-of-the-art models. However, individual models’ performances need to be considered in assigning weights to the models in the fusion.

5) The BF, though a computation-intensive method, obtained better results by searching all possible combinations of weights and choosing the one with the best results. However, in applications with fewer models (as we have in our application), the computational complexity is negligible.

6) The difference in the performances of BF and other competing methods, such as PSO, GA, and Powell’s method, is negligible.

7) The use of NLP and ML techniques will allow us to automatically acquire/retrieve and analyze the relevant posts containing citizens’ complaints about drinking water as an additional source of information. This automatic analysis will result in a significant reduction in the efforts and time spent on manual feedback through crowdsourcing techniques.

V. CONCLUSION AND FUTURE WORK

In this article, we proposed an ensemble framework for water quality analysis in social media posts. To this aim, different preprocessing, data augmentation, classification, and fusion strategies are analyzed and evaluated. Though the social media posts contain images, we focused on textual information only. Our choice of using textual information only is mainly motivated by the quality and quantity of the images associated with the posts. Overall, we used three state-of-the-art NN models individually and jointly in both naive and merit-based fusion methods. During the experiments, we observed better performance for merit-based fusion schemes, where weights were assigned to models based on their performance. This emphasizes the assumption that individual performances of the models should be considered in fusion.

In the future, we aim to incorporate the additional information available in the form of images and metadata to further enhance the performance of the framework. However, the use of the additional information, especially the images
associated with social media posts, is subject to the availability and quality of the images. To explore this aspect of the problem, we also aim to assemble a collection of social media posts containing images along with the text. We believe the additional information in the form of videos and images, for example, showing the bad color of water, will complement the textual information.

REFERENCES

[1] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and P. Halvorsen, “Social media and satellites: Disaster event detection, linking and summarization,” Multimedia Tools Appl., vol. 78, no. 3, pp. 2837–2875, 2019.

[2] J. Boddy and L. Dominelli, “Social media and social work: The challenges of a new ethical space,” Aust. Social Work, vol. 70, no. 2, pp. 172–184, 2017.

[3] M. Ofori and O. El-Gayar, “Drivers and challenges of precision agriculture: A social media perspective,” Precis. Agr., vol. 22, no. 3, pp. 1019–1044, 2021.

[4] A. Ben-David and A. M. Fernández, “Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme right political parties in Spain,” Int. J. Commun., vol. 10, p. 27, Feb. 2016.

[5] A. Matamoros-Fernández and J. Farkas, “Racism, hate speech, and social media: A systematic review and critique,” Television New Media, vol. 22, no. 2, pp. 205–224, 2021.

[6] K. Müller and C. Schwarz, “Fanning the flames of hate: Social media and hate crime,” J. Eur. Econ. Assoc., vol. 19, no. 4, pp. 2131–2167, 2021.

[7] S. B. Naeem, R. Bhatti, and A. Khan, “An exploration of how fake news is taking over social media and putting public health at risk,” Health Inf. Libraries J., vol. 38, no. 2, pp. 143–149, 2021.

[8] N. Said et al., “Natural disasters detection in social media and satellite imagery: A survey,” Multimedia Tools Appl., vol. 78, no. 22, pp. 31267–31302, 2019.

[9] A. Bodaghi and J. Oliveira, “The theater of fake news spreading, who plays which role? A study on real graphs of spreading on Twitter,” Expert Syst. Appl., vol. 189, Mar. 2022, Art. no. 116110.

[10] N. Rapousis, M. Katsarakis, and M. Papadopouli, “QoWater: A crowdsourcing approach for assessing the water quality,” in Proc. 1st ACM Int. Workshop Cyber-Phys. Syst. Smart Water Netw., 2015, pp. 1–6.

[11] M. Galvis, B. Schaeffer, R. Parmar, K. Wolfe, and J. M. Johnston (Project Environ. Protection Agency, Nat. Aeronaut. Space Admin. (NASA), Nat. Ocean. Atmosp. Admin. (NOAA), United States Geol. Surv. (USGS), Washington, DC, USA), Cyanobacteria Assessment Network (CyAN). [Online]. Available: https://www.epa.gov/water-research/cyanobacteria-assessment-network-cyan

[12] A. Molsen, M. Elshemy, and B. Zedian, “Water quality monitoring of lake burullus (Egypt) using landsat satellite images,” Environ. Sci. Pollut. Res., vol. 28, no. 13, pp. 15687–15700, 2021.

[13] J. R. Fisher, E. A. Acosta, P. J. Denny-Frank, T. Kroeger, and T. M. Boucher, “Impact of satellite imagery spatial resolution on land use classification accuracy and modeled water quality,” Remote Sens. Environ. Conserv., vol. 4, no. 2, pp. 137–149, 2018.

[14] M. Bonansea, M. Ledesma, C. Rodríguez, and L. Pinotti, “Using new remote sensing satellites for assessing water quality in a reservoir,” Hydrol. Sci. J., vol. 64, no. 1, pp. 34–44, 2019.

[15] S. Jakositz et al., “Protection through participation: Crowdsourced tap water quality monitoring for enhanced public health,” Water Res., vol. 169, Feb. 2020, Art. no. 115209.

[16] A. N. X. Mix and A. Haas, “Social media monitoring for water quality surveillance and response systems,” AWWA Water Sci., vol. 112, no. 8, p. 44, 2020.

[17] H. Zheng, Y. Hong, D. Long, and H. Jing, “Monitoring surface water quality using social media in the context of citizen science,” Hydrol. Earth Syst. Sci., vol. 21, no. 2, pp. 949–961, 2017.

[18] 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.

[19] 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, Jun. 2021, Art. no. 126814.

[20] H. Do, “Mining social media to assess public perception of water quality,” M.S. thesis, Nichols School Environ., Duke Univ., Durham, NC, USA, 2021.

[21] S. Andreadis et al., “Water-MM: Water quality in social multimedia,” in Proc. MediaEval Workshop, 2021.

[22] 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 Workshop, 2021, pp. 1–3. [Online]. Available: https://2021.multimediaeval.com/paper29.pdf

[23] M. A. Ayub, K. Ahmad, N. Ahmad, and A. Al-Fuqaha, “Nlp techniques for water quality analysis in social media content,” in Proc. MediaEval Workshop, 2021, pp. 1–3. [Online]. Available: https://2021.multimediaeval.com/paper23.pdf

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

[25] Y. Liu et al., “Roberta: A robustly optimized bert pretraining approach,” 2019, arXiv:1907.11692.

[26] K. Ahmad, K. Khan, and A. Al-Fuqaha, “Intelligent fusion of deep features for improved waste classification,” IEEE Access, vol. 8, pp. 96495–96504, 2020.

[27] Y. Del Valle, G. K. Venayagamoorthy, S. Mohagheghi, J.-C. Hernandez, and R. G. Harley, “Particle swarm optimization: Basic concepts, variants and applications in power systems,” IEEE Trans. Evol. Comput., vol. 12, no. 2, pp. 171–195, Apr. 2008.

[28] M. J. Powell, “An efficient method for finding the minimum of a function of several variables without calculating derivatives,” Comput. J., vol. 7, no. 2, pp. 155–162, 1964.

[29] S. Singer and J. Nelder, “Nelder-Mead algorithm,” Scholarpedia, vol. 4, no. 7, p. 2928, 2009.

K. Ahmad received the B.Sc. degree from the University of Engineering and Technology Peshawar, Peshawar, Pakistan, in 2021. He is currently pursuing the M.Sc. degree in information and telecommunications engineering from Jeonbuk National University, Jeonju, South Korea.

He is also working as a Research Assistant with the Robot Vision Laboratory and Intelligent Robot Research Center. His research interests include multimedia analysis, NLP, computer vision, and vision intelligence.

M. A. Ayub received the B.Sc. degree from the University of Engineering and Technology, Peshawar, Peshawar, Pakistan, in 2021. He is currently working as a Research Assistant with the Department of Computer Systems Engineering, University of Engineering and Technology. His research interests include multimedia analytics, AI, ML, and NLP.

K. Ahmad received the bachelor’s and master’s degrees from the University of Engineering and Technology Peshawar, Peshawar, Pakistan, in 2010 and 2013, respectively, and the Ph.D. degree from the University of Trento, Trento, Italy, in 2017.

He worked with the Multimedia Laboratory in DISI, University of Trento. He is currently working as a Lecturer with the Department of Computer Science, Munster Technological University, Cork, Ireland. He has also worked as a Postdoctoral Researcher with Hamad Bin Khalifa University, Doha, Qatar, and ADAPT Centre, Trinity College, Dublin, Ireland. He has authored and coauthored more than 70 journal and conference publications. His research interests include multimedia analysis, computer vision, ML, and signal processing applications in smart cities.

Dr. Ahmad is a Program Committee Member of multiple international conferences, including CBMI, ICIP, and MMSys.
J. Khan received the B.Sc. and M.Sc. degrees in computer systems engineering from the University of Engineering and Technology Peshawar, Peshawar, Pakistan, in 2010 and 2017, respectively, and the Ph.D. degree in electronics and information engineering from Korea Aerospace University, Goyang, South Korea, in 2021.

He is currently pursuing a Postdoctoral Fellow with the Department of Artificial Intelligence, AJOU University, Suwon, South Korea. His research interests include social networks analysis, AI, NLP, and recommendation systems.

Nasir Ahmad received the B.Sc. degree in electrical engineering from the University of Engineering and Technology Peshawar, Peshawar, Pakistan, in 2004, and the Ph.D. degree from Loughborough University, Loughborough, U.K., in 2011.

He is currently working as an Associate Professor with the Department of Computer Systems Engineering, University of Engineering and Technology Peshawar. His research interests include speech and video processing digital signal processing, pattern recognition, and machine learning.

A. Al-Fuqaha (Senior Member, IEEE) received the Ph.D. degree in computer engineering and networking from the University of Missouri-Kansas City, Kansas City, MO, USA, in 2004. He is currently a Professor with Hamad Bin Khalifa University, Ar-Rayyan, Qatar. His research interests include the use of machine learning in general and deep learning in particular in support of the data-driven and self-driven management of large-scale deployments of IoT and smart city infrastructure and services, wireless vehicular networks, cooperation and spectrum access etiquette in cognitive radio networks, and management and planning of software-defined networks.

Prof. Fuqaha serves on editorial boards of multiple journals, including IEEE Communications Letter and IEEE Network Magazine. He also served as the Chair, the Co-Chair, and the Technical Program Committee Member of multiple international conferences, including IEEE VTC, IEEE Globecom, IEEE ICC, and IWCMC. He is a Senior Member of ABET Program Evaluator.