Harnessing Social Media Analysis to Enhance Water Management Efficiency

Khalid K Osman[1] , **Florian Barbaro**[2] and **Andy Skumanich**[2]
[1]Stanford University
[2]Innov8ai
osmank@stanford.edu, fbarbaro@innov8ai.com, askuman@innov8ai.com

Abstract

Access to clean potable water is a fundamental requirement for human well-being. However, numerous challenges arise in water delivery and management, particularly in the context of global climate change. The increasing severity of droughts, social inequalities, and infrastructure constraints necessitates a concerted effort towards water management. Despite the rise of Social Media as a platform for societal engagement across various topics, the inclusion of water concerns in this discourse remains remarkably underutilized, undeveloped, and underanalyzed.

This paper aims to highlight the untapped opportunities offered by social media as a valuable tool for water management. We emphasize the potential of social media to serve as a versatile medium for collecting and disseminating information regarding water-related issues. By leveraging advanced AI and Machine Learning techniques, we conduct textual analysis on a comprehensive dataset of water utility tweets in the United States, extracting pertinent and meaningful insights. These insights serve as a starting point for further advancements in water management.

1 Introduction

There are multiple aspects of using social media for water management [Johns, 2014] which range from customer sentiment (such as direct feedback about water quality) to more regional or global aspects of *water anxiety*. The former can be part of a water utility's engagement with its customer base, and social media provides a useful path of underutilized communication. However, the latter aspect of general water concerns is of increasing importance given the developing global water supply issues.

In this paper, we indicate the importance of both ends of this spectrum, namely how the water utilities need to expand the immediate outreach for the benefit of their customers [Solis and Bashar, 2022], and also how high-level data analysis can show regional water aspects [Shafiee *et al.*, 2018]. We focus on the latter regional dimension in this paper, but an important point is that the full range of this spectrum needs to be monitored and addressed to improve the overall functioning of water infrastructure management for the benefit of society in challenging times and that the proper use of social media can assist in that endeavour.

A recent publication [Foundation, 2017] focused on how water utilities use social media to reach their customers and determined that the water sector was behind other industries, including the electric sector, in using social media. This study analyzed the social media profiles of utilities across the United States and showed that only a small percentage of them were using social media, and an even smaller fraction reached their customer base. The range of engagement varied widely and highlighted the challenge of using social media. We propose that this shortcoming needs to be addressed because of the potential value of social media for not just monitoring customer satisfaction but for general water management using crowd-sourcing and citizen scientist types of engagements. However, given space limitations, our discussion in this paper focuses on the more global aspect of social media as reflecting *water anxiety* or other social concerns from a more macro perspective. We find that by using a refined AI/ML approach along with interdisciplinary social science, we can extract regional expressions which can be important to the global water community and could lead to improved overall water management strategies.

The rest of the paper is organized as follows. In Section Context we introduce the context of water management for the three states studied in the article. In Section Methods we present the dataset and the main models used in this article. Then, in Section Results, we introduce our clustering and classification results. Finally, in Section Summary of Results, we summarize the results obtained and in Section Conclusion and broader perspective we discuss possible extensions of our work and its impacts.

2 Context

Water management risks are increasing from a variety of challenging aspects and they vary by specific location or by region. In particular, as representative examples, California, Florida, and Utah all face their own unique set of risks, which are still common to global risks.

In California, a principal water management risk is due to climate change. With an increasing demand for water, California is facing a dramatic drop in water availability. This is evidenced by surface water sources like reservoirs, rivers and lakes which are depleted due to extended droughts. Additionally, there is significant demand from the agricultural sector which constitutes a major part of the water use in the state. Further, groundwater resources are being over-pumped which can lead to land subsidence and reduced aquifer levels. As a result, water management has become one of the most important issues to date. Drought conservation requirements have already been implemented.

In Florida, although there is no drought as in the southwestern part of the US, one of the primary water management risks is related to saltwater intrusion from rising sea levels associated with climate change. Saltwater intrusion occurs when saltwater from the ocean enters freshwater aquifers as sea levels rise due to melting ice caps or storm surge flooding events– this can contaminate drinking water supplies as well as harm local ecosystems. Additionally, Florida faces an increased risk of flooding due to sea level rise which can inundate coastal communities and damage infrastructure associated with the water supply.

In Utah, one of the primary water management risks is drought associated with climate change combined with population growth leading to an increased demand for water resources. As decreasing precipitation rates and prolonged dry spells continue throughout much of the state's arid regions this has led to a significant decrease in available surface water sources such as rivers or lakes while also reducing groundwater supplies through over-pumping or depletion of aquifers by agricultural demands. In addition, many areas throughout Utah have already seen decreases in snowpack accumulation (similar to California) which further exacerbates this issue by reducing runoff into streams or springs that serve as major sources of drinking water supplies for many rural communities across the state.

Overall each state has its own unique set of challenges when it comes to managing their respective watersheds effectively– however, all three states are facing similar issues related primarily to climate change that could continue to accelerate existing problems if left unaddressed such as decreased availability of surface waters sources and regional droughts. Hence it will be important for all three states to take proactive steps towards addressing these issues before they become too complex to address. Multiple tools must be brought to bear to assist in solving these problems and one critical tool is the use of collective social resources - crowdsourcing. An important aspect to implement social responsibility is the use of social media. It is in this context that we take a closer look at the aspect of water awareness and how that is reflected in current social media.

3 Related Work

Already it has become common for disaster management organizations and for local utilities to use social media (e.g. Twitter and Facebook) to provide warnings. These are typically for immediate crisis conditions [Hughes and Palen, 2009; Wu and Cui, 2018]. Since access to traditional sources (e.g., TV, landlines, radio) can be constrained for individuals, emergency management organizations use social media [Hughes and Palen, 2009; Wu and Cui, 2018]. There is evidence from other studies that social media is underutilized as a key communication channel during disasters [Smith *et al.*, 2018]. Also in reverse, social media platforms have also served as a tool for the public to inform authorities and neighbours that they are safe or that the immediate danger has passed. As a specific example, Facebook provides individuals with a mode to declare they are safe using location-based data when a disaster event occurs nearby [Stephens *et al.*, 2020].

Note further that the typical orientation of social media is on immediate or imminent disasters. Existing studies have focused on individual social media behaviours during disasters [Bean *et al.*, 2021; Hughes and Palen, 2009; Jamali *et al.*, 2019], or the development of social media modes for large-scale emergency management [Houston *et al.*, 2015; Merchant and Lurie, 2020], and damage assessment using social media activity [Kryvasheyeu *et al.*, 2016; Nguyen *et al.*, 2017; Wu and Cui, 2018]. While this is important, it does not fully capture the relationship between end-users and the communication received from utility providers during crisis situations. Water providers are oriented to understanding the crisis from a local context, specifically in how everyday water infrastructure services may be impacted by a given disaster. And further, this has almost entirely been on immediate issues of a disaster, not the broader types of issues associated with climate change. It is increasingly necessary to look at broader issues, such as regional droughts, which need to be communicated to the community.

4 Methods

This section describes the overall approach of the proposed application. The details of the implementation are presented next.

4.1 Data

Dataset collection

The data consisted of water utilities' tweets in the aftermath of the COVID-19 pandemic. Examining the tweets shared by water utilities during this time is crucial in understanding how they disseminate information and address public concerns regarding the virus's potential presence in the water. These tweets provide valuable insights into the communication strategies employed by utilities to mitigate fears and ensure public confidence in tap water safety.

Tweets along with engagement data (likes, retweets) were extracted directly from water utilities' Twitter pages utilizing a Twitter Developer Application Programming Interface (API) [Makice, 2009]. The water utilities examined in this study were selected based on whether their county met the US Center for Disease Control (CDC) hotspot classification [Oster *et al.*, 2020] at least one time between March 2020 and October 2020 for three states, that is to say, California, Florida and Utah. It was critical that we captured data early in the pandemic so that we are analyzing crisis response immediately or shortly after the pandemic began. Water utilities within the counties were then identified by manually searching county websites.

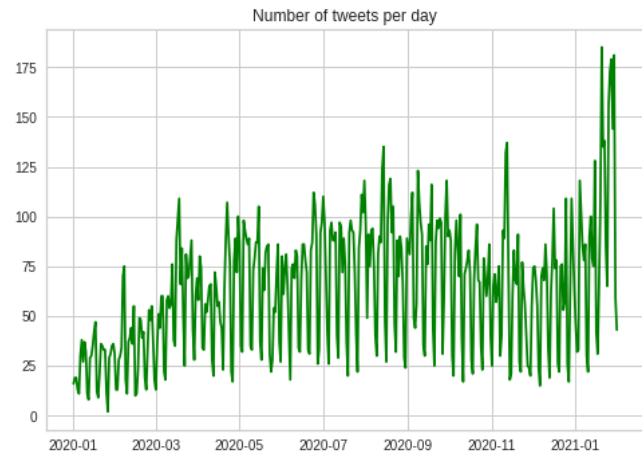

Figure 1: Number of Tweets per day for the three states.

Over 25000 tweets were extracted from 40 water utilities from January 01, 2020 (World Health Organization declared COVID-19 a public health emergency on January 30, 2020 [Cucinotta and Vanelli, 2020]) to January 30, 2021. Table 1 shows the distribution of tweets and utilities by state. The database is heterogeneous containing mostly Tweets from counties in Florida and shows the difference in the propagation of Covid in each state.

Each Twitter account for the utilities was manually verified by visiting each profile as many counties had multiple utilities. Utilities without a Twitter presence were excluded from the study. Additionally, some of the tweets regarding water utilities were posted by town, city, or county-level Twitter accounts rather than the utilities themselves. This was particularly true for smaller counties where everything was managed by one entity. These tweets remained in the study and were treated with the same veracity as those from utilities.

Table 1: Distribution of tweets and utilities by state

State	Nb. of tweets	Nb. of utilities
California	5521	13
Florida	17957	21
Utah	1657	6
Total	25135	40

Figure 1 shows the Number of tweets per day. We can see a clear increase in the number of tweets over time. Figure 2 shows the number of new Covid cases per day in the US. It is interesting to note the similarity of the peaks in the number of tweets and the number of new Covid cases.

Data preprocessing

We used two distinct strategies for text processing depending on the algorithm used. For BERT, we kept the text as it was. However, for other analyses, we lowercase the text to apply a traditional pre-processing by removing stop-words and digits, and lemmatisation [Barbaro, 2022]. Then we kept the N-grams from $N = 1$ to $N = 4$ and removed the rare ones (appearing in less than 1% of documents) to obtain 8242 N-grams. Finally, we used the term frequency–inverse document frequency (TF-IDF) to weigh the N-grams.

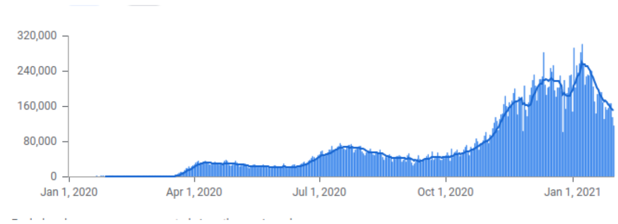

Figure 2: Number of new Covid cases per day in US.

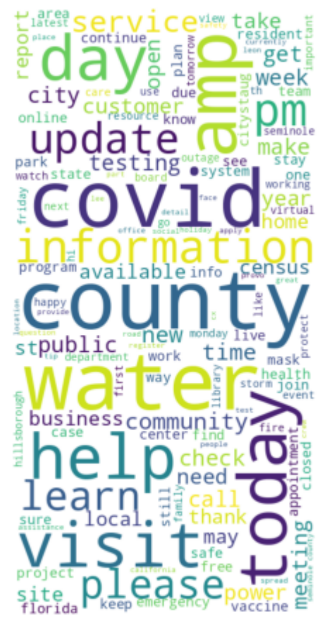

Figure 3: Word clouds of Water Management for California, Florida and Utah.

Word clouds are a simple but effective tool for text visualization. They are created by collecting words in a corpus and presenting them in different sizes. The larger and bolder an N-gram appears the greater its TF-IDF weight in the corpus. Figure 3 illustrates the world cloud. This word cloud presents the dominant topics of water management challenges, including terms representing the daily tasks of water management utilities such as information, help and water. Furthermore, what is interesting to note in this corpus is the strong presence of the term-related Covid and the fact that utilities had to reassure the population about the quality of the water.

4.2 Models

Spherical k-means (Sk-means)

Many models are poorly suited to high-dimensional data, including those derived from the vector representation of text [Barbaro and Rossi, 2021]. When the data is directional [Mardia and Jupp, 2009], i.e. when it is their correlation rather than their Euclidean distance that matters, Gaussian-type models are even less suitable. For such data, it is natural to carry out a normalisation that places them on the unity sphere. Then one can use clustering techniques that address specifically the fact the data are spherical, such as Spherical k-means [Dhillon *et al.*, 2002].

The Spherical k-means algorithm (Sk-means), originally proposed in [Dhillon *et al.*, 2002], is a simple adaptation of the k-means algorithm to the cosine dissimilarity. Let us consider a collection of N observations $\boldsymbol{X} = (\boldsymbol{x}_i)_{1\leq i\leq N}$ on the hypersphere $\mathbb{S}^{d-1}$. Given a number of clusters K, Sk-means tries to find a set of K prototypes $(\boldsymbol{\mu}_k)_{1\leq k\leq K}$ in $\mathbb{S}^{d-1}$ and a clustering/membership $\boldsymbol{Z} = (z_i)_{1\leq i\leq N}$, that assigns $\boldsymbol{x}_i$ to cluster $z_i \in \{1, \dots, K\}$ such that the coherence

$$\mathcal{Q}((\boldsymbol{\mu}_k)_{1\leq k\leq K}, (z_i)_{1\leq i\leq N}) = \sum_{i=1}^{N} \boldsymbol{\mu}_{k_i}^T \boldsymbol{x}_i \tag{1}$$

is maximal.

Several methods have been proposed to maximize the coherence (see e.g. [Hornik *et al.*, 2012]).

In the following experience, we used the implementation of [Kim *et al.*, 2020], which proposes a fast initialization and enforces sparsity on the centroid vectors by using a data-driven threshold that is capable of dynamically adjusting its value depending on the clusters.

BERT

BERT (Bidirectional Encoder Representations from Transformers) is an embedding layer designed to train deep bidirectional representations from unlabeled texts by jointly conditioning on both left and right context in all layers. The BERT model consists of two steps: pre-training and fine-tuning [Devlin *et al.*, 2019]. During pre-training, the model is trained on an unlabeled large corpus. For fine-tuning, the model is initialized with the pre-trained parameters and all the parameters are fine-tuned using labelled data for specific tasks. BERT's model architecture is a multi-layer bidirectional Transformer encoder [Devlin *et al.*, 2019] based on the original implementation described in [Vaswani *et al.*, 2017]. This kind of encoder is composed of a stack of N = 6 identical layers. Each of these layers has two sub-layers. The first one is a multi-head self-attention mechanism, and the second one is a simple position-wise fully connected feedforward network. It employs a residual connection [He *et al.*, 2015] around both sub-layers, followed by a layer normalization [Vaswani *et al.*, 2017].

In the following experience, we used the model pre-trained in [Hayawi *et al.*, 2022] to tackle the disinformation about Covid based on bert-large-uncased.

5 Results

This section describes the use of Sk-means for clustering to analyze the data and then show a binary classification using BERT for information credibility.

5.1 Clustering using Sk-means

First of all, the main problem when using clustering models is selecting the number of clusters. To do so, we used the Elbow method and got K=8.

Then we applied the Spherical k-means to the water management corpus focusing on three states: Florida, California and Utah. Table 2 shows the distribution of reports by cluster obtained. The distribution of Tweets in each cluster is quite heterogeneous with two important clusters, 2 and 8, a minor cluster, 3, and the rest of the clusters which are quite homogeneous between them.

Table 2: Distribution of tweets by cluster

Clusters	Nb. of News
1	2964
2	4965
3	879
4	2604
5	2534
6	2446
7	3500
8	5243

Thanks to the implementation of sk-kmeans from [Kim *et al.*, 2020] which enforces the sparsity on the centroid vectors, we are able to retrieve the most important N-grams by cluster. The table 3 shows the top 10 N-grams for each cluster.

It is interesting to note that these N-grams defining clusters express notions that are essential to the risks faced by Water Management utilities. Cluster 1 is related to water supply and related environmental issues. Cluster 2 pertains to city commissions or other forms of local government meetings. It appears that these entities are convening or discussing various issues related to their communities. Clusters 3, 4, 7 and 8 discussed Covid related efforts deployed by their utilities or counties to stop the spread of the illness and keep people up to date on the situation. Cluster 5 is related to power outages which could disrupt daily life activities and can cause problems with Water Management. People are asked to keep abreast of any news about the restoration efforts of the utilities concerned. The focus for Cluster 6 appears centred around census initiatives, suggesting citizens have been encouraged by their governments/local leaders to take part in this year's count; they're also being implored online using hashtags (#MakeItCount). As well, respondents might be directed towards specific websites where they can fill out details digitally instead of having face–face interactions during door–knocking campaigns traditionally employed by enumerators collecting data manually across neighbourhoods nationwide annually every decade since 1790.

Figure 4 shows the distribution of state by cluster. As we can observe, all clusters contain each state, however with notable differences. California, as mentioned earlier, is more concerned with water supply as shown in Cluster 1 and therefore needs to communicate on the many issues that may arise with Cluster 5. Florida, which is in the majority in clusters 3, 4 and 7, is focusing on monitoring the evolution of Covid and informing the population. Utah focuses on the responsibility of informing its citizens about decisions made by public authorities as well as inviting its citizens to participate in the state's public life, as evidenced by clusters 2 and 6. Finally,

Table 3: Top 10 words by cluster for Sk-means

Clusters	Top words
cluster 1	water supply, gallon water, water wednesday, water news, water project, state water project, water education, save water, emergency notifications, notifications
cluster 2	city commission, citystaugtbr, citystaug citystaugtbr, convene, city commission meeting, board director meeting, director meeting, meeting packet, live online, watch online
cluster 3	latest report, latest report visit, view latest report, view latest report visit, positivity rate, nonresident, rate positive, covid test performed, covid test performed highlandscounty, dashboard report
cluster 4	local covid case, update local covid case, local covid case currently, currently hospitalized death highlandscounty, death highlandscounty, hospitalized death highlandscounty, covid testing site, open patient, case reported past, case reported past day
cluster 5	power restored, restore power, experiencing power, experiencing power outage, restoration time, estimated restoration, estimated restoration time, power please, power please check, check main
cluster 6	take census, makeitcount, census today, fill census, census online, respond census, makepinellascount, citystaug census, complete census online, respond online
cluster 7	board county commissioner, covid statistic, county covid statistic, seminole county covid, seminole county covid statistic, learn seminole, learn seminole county, learn seminole county covid, county government, commissioner meeting
cluster 8	stop spread, help slow, help slow spread, wear mask, help slow spread covid, washing hand, slow spread covid, slow spread, wearing mask, face mask

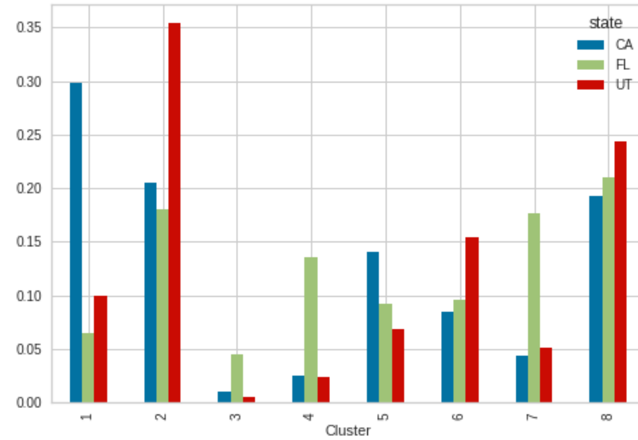

Figure 4: Distribution of State by cluster.

it should be noted that these three states advise their citizens to follow the health procedures to protect themselves from Covid despite different political opinions as shown by Cluster 8.

We note that despite the mix of several states with different water management policies, essential information for each state is not diluted and enables localized aspects to be obtained. This is the type of information which could be used by governments and regional authorities to direct messaging or resources to engage with the appropriate community. Given the limitation of resources, correct targeting can be essential. The fact that no *water anxiety* is showing up in the social media for UT is in itself a point of note. In UT there was no reflection of *water anxiety* even though there has been a significant drought in the just prior 2018-2019 time frame along with other water supply issues. Importantly, this is a major developing issue in that experts warn that the Great Salt Lake is set to disappear in the next five years [Abbott *et al.*, 2023]. Abbott said “The grim climate reality already unfolding in the Great Salt Lake, is a microcosm of what is happening or is set to happen around the world on a warming planet. “This is a bellwether for what’s going on in the larger river basins,” he said. “We need to lay out some very clear language about where we’re headed.” The main solution that the authors propose includes public awareness of conservation and the key to this type of mobilization is through social media. This is exactly the type of hybrid of social media and social sciences which needs to be accelerated. Our analysis of the UT social media shows the need for more direct social media action to develop social awareness.

Some of the examples of the tweets are as follows and show the information neutrality associated with Covid (no misinformation) in Figure 5.

5.2 Binary classification using BERT for information credibility

A critical aspect of social media is the level of credibility [Barbaro and Skumanich, 2023] and this is a separate focus of our paper. To do so, we first fine-tuned BERT on a labelled database and then applied it to our data.

State: Florida
Tweet: Wearing a mask is a sign of respect for our neighbours and helps you to stop the spread of COVID-19. Learn more at Let's work together to prove that HillsboroughCares.

State: California
Tweet: COVID-19 will not impact the quality or supply of your tap water. You can continue to enjoy clean and reliable water straight from the tap.

Figure 5: Example of tweets that are informationally neutral associated with Covid

We used BERT fine-tuned on the database from [Hayawi *et al.*, 2022]. They performed manual annotation of tweets about vaccine misinformation which was controlled by medical experts. Consequently, a total of 15.073 tweets were labelled, 5751 of which were misinformation and 9322 general vaccine-related tweets. Moreover, we followed the same procedure. We first trained and validated 75% of the dataset and then evaluated the remaining 25% of the dataset. BERT was trained for three iterations with a 20% validation set taken from a subset of the training set. Finally, we obtained the same results as presented in [Hayawi *et al.*, 2022] with an accuracy and an F1-score of 98% on the test set.

We applied it to our database. As we fine-tuned for covid misinformation, we kept only Covid tweets related, that is to say, 4344 tweets. Then, we evaluated these tweets regarding misinformation. In the cases of the water utilities' social media, there was no detected misinformation. Instead, the content provided useful and relevant information. Note that this finding is important in that it demonstrates that some sources of Covid information can be socially augmenting. The connection with water utilities is that the water sources can be post-monitored for viruses and can provide important metrics regarding infections and local or regional levels (which manifest in the tested wastewater). In principle, this information could be used as a local health alert.

6 Summary of Results

The key point of note is that key information and crisis awareness aspects for a given region for a given time period, are manifested by the AI analysis in a detectable way. The AI analysis of social media provides a critical tool for determining the current sentiment, or the level of issue awareness, and also the trending concerns, which may not be reflected by standard modes.

The clustering shows significant and meaningful results. The California clustering has a dominant cluster associated with *water anxiety*, whereas Florida does not. Instead, the Florida clusters are associated with *Covid anxiety* which is an adjacent social issue. In the case of Utah, there is no *water anxiety* but reduced *Covid anxiety*. Utah does not have the same level of drought as California, however, it is developing and would become a major issue. It is important to note that the Utah data did not show *water anxiety*. This could indicate the need for improved social communication about developing issues.

In addition, we observed that the information published by the public services was of high quality and could be used for local health alerts regardless of the political views pursued by the different states.

Because the key "crisis awareness" points can be detected, the Utah government and the US federal government can target a regional messaging system to increase awareness of the developing issues with the water supply (i.e. that the Great Salt Lake will be in dire straits in 5 years). By using AI for active monitoring of the clusters, it would be possible to determine the level of public awareness development. In this way, AI can be a key tool to help drive the targeted messaging and evaluate its effectiveness. The modes to get public sentiment can often be very challenging and this tool gives policymakers a definite method for assessment.

7 Conclusion and broader perspective

The intent of this paper is meant to be indicative and not definitive, with the purpose of showing the value of a cross-cutting combination of social media, social sciences, and AI analytics to address pressing social considerations and critical problems which will be affecting the global communities with a very high level of urgency.

In this article, we observed that with even a limited set of social media data, key societal and water issues can be extracted with the appropriate AI/ML analysis along with interdisciplinary social science. This conclusion is important as a starting point given how much in retrograde the water utilities and water management are with social media and the developing urgency of water issues. The data we analyze shows how critical the use of social media will be not just at the customer relations level (not developed in detail in this paper), but at a global level looking at the aspects of, for instance, *water anxiety*. This type of analysis can allow governments and policy organizations to better detect and target optimization efforts. In essence, the paper provides a methodology which can be expanded upon. More robust conclusions can be derived with larger or more targeted data sets. Note that in this case, there is an opportunity to use water utilities as viable sources for important social information which can have a broader impact. In particular, because water utilities can sample wastewater, they can be partners in identifying viral spread or occurrence. As the utilities are encouraged to expand their online messaging, they could provide information which can be helpful to larger agencies looking at viral spread. In addition, the appropriate use of social media can include crowd-sourcing and citizen scientist inclusion which is an invaluable asset for managing critical social challenges such as water distribution. This would be a starting point for socially inclusive data which can be very important for optimal management of a diminishing or valuable critical human needs resource. The overarching conclusion is that social media is now essential for addressing critical issues and that analytics should be exploited to most effectively evaluate where and how to use social media.

References

[Abbott *et al.*, 2023] Benjamin Abbott, Bonnie Baxter, Karoline Busche, Lynn Freitas, Rebecca Frei, Teresa Gomez, Mary Karren, Rachel Buck, Joseph Price, Sara Frutos, Robert Sowby, Janice Brahney, Bryan Hopkins, Matthew Bekker, Jeremy Bekker, Russell Rader, Brian Brown, Mary Proteau, Greg Carling, and Patrick Belmont. Emergency measures needed to rescue great salt lake from ongoing collapse. 01 2023.

[Barbaro and Rossi, 2021] Florian Barbaro and Fabrice Rossi. Sparse mixture of von mises-fisher distribution. In *ESANN 2021 proceedings*. Ciaco - i6doc.com, 2021.

[Barbaro and Skumanich, 2023] Florian Barbaro and Andy Skumanich. Addressing socially destructive disinformation on the web with advanced ai tools: Russia as a case study. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 204–207, New York, NY, USA, 2023. Association for Computing Machinery.

[Barbaro, 2022] Florian Barbaro. *Analyse exploratoire et classification de textes*. Theses, Paris 1 - Panthéon-Sorbonne, June 2022.

[Bean *et al.*, 2021] Hamilton Bean, Ana Maria Cruz, Mika Shimizu, Keri K Stephens, Matthew McGlone, and Sharon Strover. Mobile alert and warning in the united states and japan: Confronting the challenges of international harmonization. *International Journal of Disaster Risk Science*, 12(6):928–934, 2021.

[Cucinotta and Vanelli, 2020] Domenico Cucinotta and Maurizio Vanelli. Who declares covid-19 a pandemic. *Acta Biomedica Atenei Parmensis*, 91(1):157–160, Mar. 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the Ass. for Comp. Ling.: Hum. Lang. Technologies*, pages 4171–4186, June 2019.

[Dhillon *et al.*, 2002] I.S. Dhillon, Yuqiang Guan, and J. Kogan. Iterative clustering of high dimensional text data augmented by local search. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 131–138, 2002.

[Foundation, 2017] The Water Research Foundation. Social media for water utilities. 2017.

[Hayawi *et al.*, 2022] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public Health*, 203:23–30, February 2022. Funding Information: This work was supported by Zayed University under the research grant RIF R20132 . Publisher Copyright: © 2021 The Royal Society for Public Health.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[Hornik *et al.*, 2012] Kurt Hornik, Ingo Feinerer, Martin Kober, and Christian Buchta. Spherical k-means clustering. *Journal of Statistical Software, Articles*, 50(10):1–22, 2012.

[Houston *et al.*, 2015] J. Brian Houston, Joshua Hawthorne, Mildred F. Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R. Halliwell, Sarah E. Turner McGowen, Rachel Davis, Shivani Vaid, Jonathan A. McElderry, and Stanford A. Griffith. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*, 39(1):1–22, 2015.

[Hughes and Palen, 2009] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3-4):248–260, 2009.

[Jamali *et al.*, 2019] Mehdi Jamali, Ali Nejat, Souparno Ghosh, Fang Jin, and Guofeng Cao. Social media data and post-disaster recovery. *International Journal of Information Management*, 44:25–37, 2019.

[Johns, 2014] Raechel Johns. Community change: Water management through the use of social media, the case of australia's murray-darling basin. *Public Relations Review*, 40(5):865–867, 2014.

[Kim *et al.*, 2020] Hyunjoong Kim, Han Kyul Kim, and Sungzoon Cho. Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling. *Expert Systems with Applications*, 150:113288, 2020.

[Kryvasheyeu *et al.*, 2016] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science Advances*, 2(3):e1500779, 2016.

[Makice, 2009] Kevin Makice. *Twitter API: Up and Running Learn How to Build Applications with the Twitter API*. O'Reilly Media, Inc., 1st edition, 2009.

[Mardia and Jupp, 2009] K.V. Mardia and P.E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 2009.

[Merchant and Lurie, 2020] Raina M. Merchant and Nicole Lurie. Social Media and Emergency Preparedness in Response to Novel Coronavirus. *JAMA*, 323(20):2011–2012, 05 2020.

[Nguyen *et al.*, 2017] Dat T. Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. Damage assessment from social media imagery data during disasters. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 569–576, 2017.

[Oster *et al.*, 2020] Alexandra M Oster, Gloria J Kang, Amy E Cha, Vladislav Beresovsky, Charles E Rose, Gabriel Rainisch, Laura Porter, Eduardo E Valverde, Elisha B Peterson, Anne K Driscoll, et al. Trends in number and distribution of covid-19 hotspot counties—united states, march 8–july 15, 2020. *Morbidity and Mortality Weekly Report*, 69(33):1127, 2020.

[Shafiee *et al.*, 2018] M. Ehsan Shafiee, Zachary Barker, and Amin Rasekh. Enhancing water system models by integrating big data. *Sustainable Cities and Society*, 37:485–491, 2018.

[Smith *et al.*, 2018] W. R. Smith, K. K. Stephens, B. R. Robertson, J. Li, and D. Murthy. Social media in citizen-led disaster response: Rescuer roles, coordination challenges, and untapped potential. *Proceedings of the ... International ISCRAM Conference*, 2018.

[Solis and Bashar, 2022] Miriam Solis and Samira Binte Bashar. Social equity implications of advanced water metering infrastructure. *Utilities Policy*, 79:101430, 2022.

[Stephens *et al.*, 2020] Keri K. Stephens, Brett W. Robertson, and Dhiraj Murthy. Throw me a lifeline: Articulating mobile social network dispersion and the social construction of risk in rescue communication. *Mobile Media & Communication*, 8(2):149–169, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[Wu and Cui, 2018] Desheng Wu and Yiwen Cui. Disaster early warning and damage assessment analysis using social media data and geo-location information. *Decision Support Systems*, 111:48–59, 2018.