9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

The potential for social media analysis to assess and optimize water management

Water Management, water anxiety, Water utilities, Spherical k-means, Clustering

Extended Abstract

An essential need for human well-being is access to adequate potable water. There are many challenges developing in the delivery and management of water, especially in the face of global climate change. With increasing severities of drought, social inequities, and infrastructure challenges, the need for concerted water management has become essential. Social and community engagement is now necessary. One of the key tools for societal engagement is the use of Social Media, and water concerns must certainly become part of this conversation. Instead, this engagement mode is underutilized and not adequately developed or analyzed.

There are multiple aspects of using social media for water management 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 (although underutilized) path of communication. However, the latter aspect of regional water concerns is of increasing importance given the developing global water supply issues. We focus on the latter.

The water utilities need to expand the immediate outreach for the benefit of their customers. In this paper, we explain how high-level data analysis can show regional water aspects which can be directed to assist the communities. The social media-derived data can be monitored and addressed to improve the overall functioning of water infrastructure management for the benefit of society in challenging times. We suggest that the proper use of social media can assist in that endeavour.

The data consisted of water utilities' tweets in the aftermath of the COVID-19 pandemic. Tweets along with engagement data (likes, retweets) were extracted directly from water utilities' Twitter pages utilizing a Twitter Developer Application Programming Interface (API). The water utilities examined in this study were selected based on whether their county met the US Center for Disease Control (CDC) hotspot classification [2] 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. 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, 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.

We performed a qualitative analysis based on a Spherical k-means algorithm to determine the main topics of disinformation and to discover the key trends with the help of the ELbow method to select the number of cluster, $k = 8$. We used the implementation of [1], 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.

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. Thanks to the implementation of sk-kmeans from [1], we are able to retrieve the most important N-grams by cluster.

The table 3 shows the top 5 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. 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). Figure 2 shows the strength of a cluster for the 3 states. As we can observe, all clusters occure in 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 the strongest 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, it should be noted that these three states are comparable in regards to the health procedures for Covid independent of different political orientations, as shown by Cluster 8.

In conclusion, 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. The clustering shows significant and meaningful results.

References

[1] Hyunjoong Kim, Han Kyul Kim, and Sungzoon Cho. "Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling". In: *Expert Systems with Applications* 150 (2020), p. 113288. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.113288. URL: http://www.sciencedirect.com/science/article/pii/S0957417420301135.

[2] Alexandra M Oster et al. "Trends in number and distribution of COVID-19 hotspot counties—United States, March 8–July 15, 2020". In: *Morbidity and Mortality Weekly Report* 69.33 (2020), p. 1127. DOI: 10.15585/mmwr.mm6933e2.

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

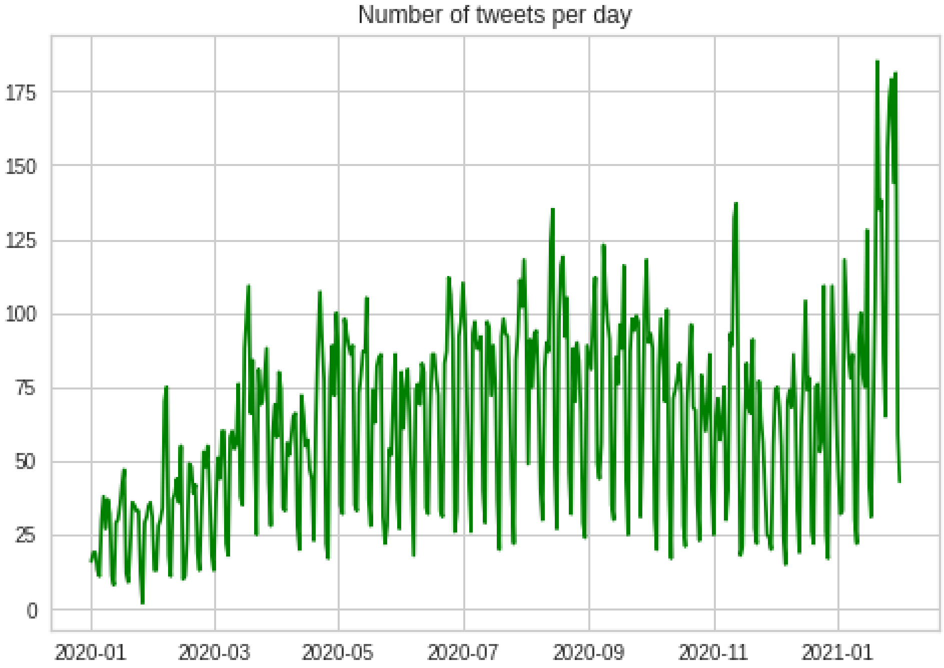

Figure 1: Number of Tweets per day for the three states.

Table 2: Distribution of tweets by cluster

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

Table 3: Top 5 words by cluster for Sk-means

Clusters	Top words
cluster 1	water supply, water news, water education, save water, emergency notifications
cluster 2	city commission, citystaugtbr, citystaug citystaugtbr, convene, city commission meeting
cluster 3	latest report, latest report visit, view latest report, view latest report visit, positivity rate
cluster 4	local covid case, update local covid case, local covid case currently, currently hospitalized death highlandscounty, death highlandscounty
cluster 5	power restored, restore power, experiencing power, experiencing power outage, restoration time
cluster 6	take census, makeitcount, census today, fill census, census online
cluster 7	board county commissioner, covid statistic, county covid statistic, seminole county covid, seminole county covid statistic
cluster 8	stop spread, help slow, help slow spread, wear mask, help slow spread covid

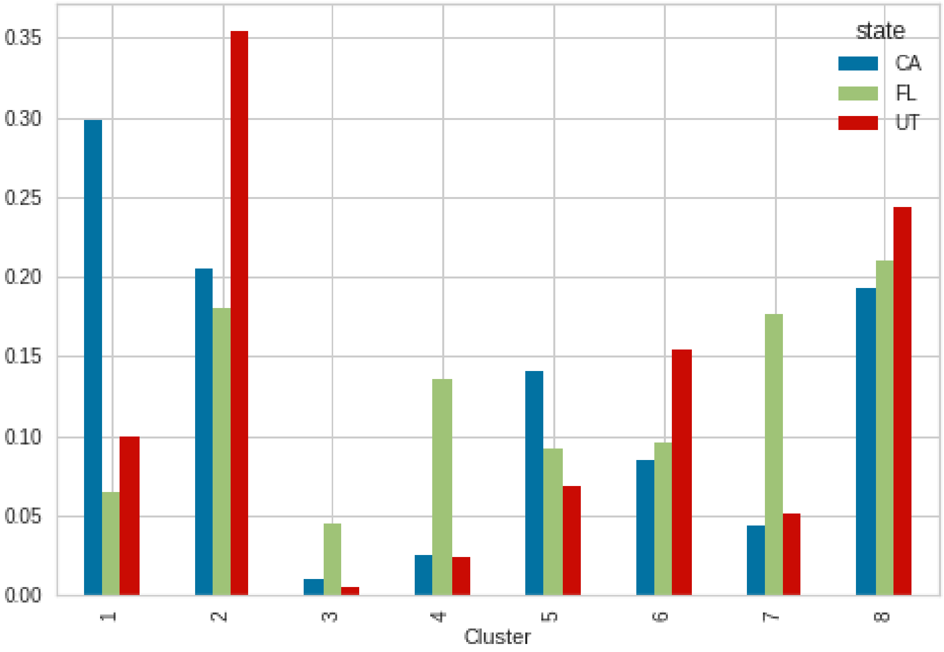

Figure 2: Distribution of State by cluster.