Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

SocialDrought: A Social and News Media Driven Dataset and Analytical Platform towards Understanding Societal Impact of Drought

Lanyu Shang[1], Bozhang Chen[1], Anav Vora[2], Yang Zhang[1], Ximing Cai[2], Dong Wang[1]

[1]School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA
[2]Civil and Environmental Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA
{lshang3, bozhang4, amvora3, yzhangnd, xmcai, dwang24}@illinois.edu

Abstract

Drought poses significant challenges to sustainability across various sectors in our society, leading to substantial consequences on agriculture, environments, ecosystems, public health, and socioeconomic stability. While prior work has studied the impacts of drought using professionally measured data sources, the societal perspectives of drought impacts remain largely under-explored. In this work, we present SocialDrought, a novel and comprehensive dataset to facilitate research on the societal impacts of drought. In particular, SocialDrought consists of three major components: 1) over 1.5 million *social media posts*, 2) over 1,400 *news articles* collected and verified by domain experts, and 3) over 31,000 *meteorological records* from the U.S. Drought Monitor about drought severity. In addition, we also introduce an online analytical platform that enables interactive and real-time data exploration to gain timely insights into the societal impacts of drought. Our interdisciplinary dataset integrates both conventional meteorological data and unconventional social and news media data to provide a holistic understanding of drought impacts. SocialDrought opens new opportunities to study the societal impacts of drought through the lens of social and news media.

Introduction

In recent years, drought has posed significant challenges to sustainability across various sectors in our communities, including agriculture (e.g., crop failures, food shortages), water resources (e.g., limited water supply, reduced water quality), and ecosystems (e.g., biodiversity decline, habitats loss). These challenges have led to non-negligible consequences that not only affect the immediate environment but also extend to broader socioeconomic stability. For example, the California drought caused an estimated $24 billion loss of agricultural revenue in 2022 (Medellin-Azuara et al. 2022). In 2023, drought and heat triggered the worst wildfire season on record in Canada which has burned more than 19.5 million acres and put more than 87 million people in North America at risk for poor air quality (Luo et al. 2024). A substantial amount of interdisciplinary efforts (e.g., Hydrology, Meteorology, Ecology, and Environmental Science) have been made to study the complexities of drought and its profound effects on both natural ecosystems and economic infrastructures (Lindersson et al. 2020). However, the societal impacts of drought among community stakeholders, such as physical and mental health issues, income fluctuations, increasing living expenses, and food security concerns, remain largely under-explored (Edwards, Gray, and Hunter 2019). In this paper, we present a comprehensive social and news media dataset along with an analytical platform to facilitate research towards understanding the societal impact of drought.

Existing studies on the societal impact of drought have mainly focused on leveraging professionally collected data, such as hydrological and meteorological records (Dantas, da Silva, and Santos 2020), economic statistics (Becker and Sparks 2020), surveys/interviews (Edwards, Gray, and Hunter 2019), and literature reviews (Lester, Flatau, and Kyron 2022), to analyze the societal impact posed by drought. However, the collection of such data often requires professional measurements or manual effort, which is both labor-intensive and time-consuming (Shang et al. 2022b). To address such a limitation, a few recent works (Smith et al. 2020; Lee et al. 2022) have started to explore publicly accessible data resources (e.g., social and online news media) to complement traditional data sources. However, these solutions either primarily focus on a specific data source (e.g., news media (Lee et al. 2022)) or only collect the data over a limited amount of time (e.g., 1-2 years (Smith et al. 2020)), making them insufficient to capture the comprehensive and long-term societal impact of drought. More importantly, as drought becomes increasingly severe and widespread (Williams, Cook, and Smerdon 2022), there is an urgent need to develop more dynamic and scalable approaches that can integrate a variety of data sources over extended periods, thereby providing a more holistic understanding of drought societal impact.

Motivated by the above knowledge gap, this paper introduces a novel and comprehensive *SocialDrought* dataset to study the societal impact of drought by including not only meteorological statistics but also social media data over more than a decade (from 2012 to 2023) and news media data from diverse sources. In addition, we also develop an online analytical platform[1] that can perform real-

[1]https://droughtweb.web.illinois.edu/

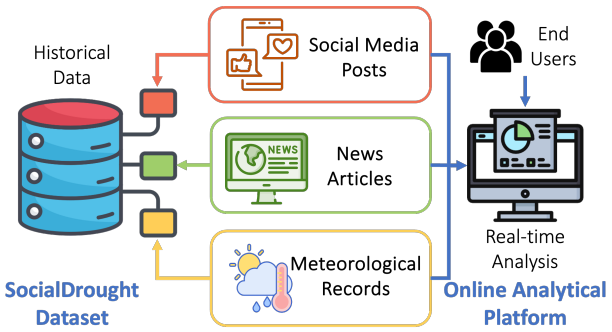


Figure 1: SocialDrought Dataset and Online Platform

time data analysis to provide timely insights from the social and news media data about a drought event/topic. The collection of this dataset and the development of the online analytical platform involve joint efforts from an interdisciplinary team of researchers in information and computer science and civil and environmental engineering, providing in-the-field know-how and domain expertise to the project's success. Figure 1 shows an overview of the SocialDrought dataset and the online analytical platform. In particular, the SocialDrought dataset consists of three parts: 1) *social media posts* that are collected using a list of keywords compiled by domain experts, 2) *news articles* that are collected daily from diverse news publishers and manually verified by domain experts, and 3) *meteorological records* that are obtained from U.S. Drought Monitor (USDM)[2] a comprehensive resource with weekly updates on drought conditions across the United States. To facilitate interactive data exploration and analysis for end users, we develop an online analytical platform that not only summarizes the highlights of the collected SocialDrought dataset but also enables end users to customize their queries and gain insights into the societal impact of drought.

The SocialDrought dataset opens up many opportunities for interdisciplinary research, improving policy-making, drought preparedness, and mitigation strategies, as well as enhancing our understanding of the social dimensions of environmental challenges. For example, historical social media posts and news articles can be leveraged to study the evolution of public opinions, concerns, and emotions at different stages of a drought event. By correlating the textual data with the meteorological statistics, researchers can analyze how public discourse and sentiment are impacted by objective drought conditions. To the best of our knowledge, SocialDrought is the first large-scale public dataset that contains 1,636,199 social media posts, 1,482 news articles, and 31,977 meteorological records. We envision the richness and versatility of the SocialDrought dataset and the online analytical platform significantly contributing to the research communities in response to the climate change challenges.

[2]https://droughtmonitor.unl.edu/

Related Work

Drought Impact

In recent years, the impacts of drought have attracted much attention across various scientific communities and government agencies, such as environmental science (Luo et al. 2024), agricultural management (Medellin-Azuara et al. 2022), public health (Salvador et al. 2020), and socioeconomics (Edwards, Gray, and Hunter 2019). Conventional methods for drought monitoring and impact assessment often leverage satellite imagery, climatic data analysis, and hydrological models to predict drought patterns, assess water resource availability, and understand ecological and socioeconomic impacts (Lindersson et al. 2020). More recently, social media and news media have been increasingly utilized as informational sources to obtain timely observations from massive social media users and professional journalists in estimating the impact of drought, especially for its impact on human society (Zhang et al. 2021a; Shang et al. 2022a). For example, MARIC (2022) developed a text classification model to categorize drought-related tweets into predefined categories to study the trends of drought impact. Lee et al. (2022) proposed a social-economic drought index based on Internet news to estimate the socioeconomic behavior and response during drought in South Korea. In this paper, we present an interdisciplinary dataset along with an online analytical platform that is dedicated to exploring drought-related social and news media information and understanding the impact of drought on our society.

Drought-related Social and News Media Datasets

Traditional drought-related datasets primarily focus on index-based statistics in related disciplines, such as meteorology (Spinoni et al. 2019), hydrology (Lai et al. 2019), and agriculture (Parsons et al. 2019). With the advances of digital devices and the ubiquity of network connections, social media and online news media emerge as rich resources in exploring and characterizing the impact of drought in human society. A few recent datasets have been collected to study social and/or news media data in the context of drought and its impact. We summarize the key characteristics of existing drought-related social and news media datasets in Table 1. In particular, Smith et al. (2020) collected a small set of 18,914 tweets from 2017 and 2018, along with news stories and drought-related indices, for the analysis of unusual drought events in the U.S. Lee et al. (2022) constructed a dataset with primarily historical news articles and observation-based drought indices to assess the socioeconomic impact of drought in South Korea. In contrast, SocialDrought is a large-scale interdisciplinary dataset that consists of 1,636,199 social media posts and 1,482 expert-verified news articles, in addition to the meteorological records of drought conditions. More importantly, our analytical platform also offers interactive opportunities (Figure 2b) to researchers and community stakeholders to search/analyze real-time drought-related social and news media data and share/contribute their own opinions/perceptions about the societal impacts of drought.

Dataset	Collection Period	Social Media	News Media	Meteorology	Geographical Area	Online Platform
DIR (Smith et al. 2020)	2017 – 2018 (2 years)	✓	✓	✓	48 Contiguous U.S. States & Washington D.C.	✗
SEDI (Lee et al. 2022)	2013 – 2017 (5 years)	✗	✓	✓	South Korea	✗
SocialDrought (Ours)	**2012 – 2023 (12 years)**	✓	✓	✓	**50 Contiguous and Non-contiguous U.S. States & Washington D.C.**	✓

Table 1: Comparison with Existing Datasets

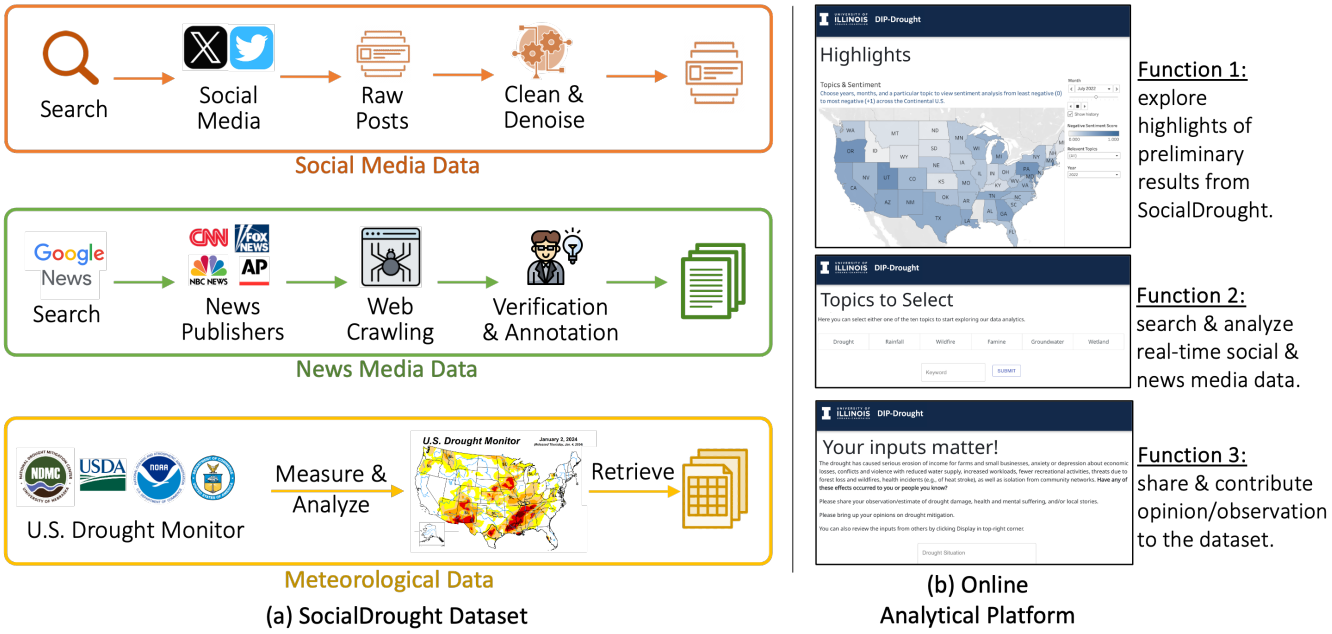


Figure 2: Overview of the SocialDrought Dataset and the Online Analytical Platform

Data Collection

Overview of SocialDrought

The SocialDrought dataset aims to leverage social and online news media to capture comprehensive and long-term information about the societal impact of drought. In particular, the SocialDrought dataset consists of three main components: 1) a set of *social media posts* that are posted by online users, 2) a set of *news articles* that are published by professional online news publishers, 3) a series of *meteorology* data that are measured and recorded by domain experts. An overview of the SocialDrought dataset and its collection process is shown in Figure 2a. We elaborate on the details of the data collection process for each component in SocialDrought below. We further analyze the data collected from these diverse sources and investigate their inter-relation towards understanding the societal impacts of drought in the Preliminary Analysis section.

Social Media Data

We choose Twitter (also known as "X") as our primary source to collect drought-related social media posts. Twitter, initially launched in 2006, is one of the mainstream social media platforms in the United States and has been adopted in diverse research domains for analyzing public opinions and activities (Zhang et al. 2021b). We leverage the *full-archive search* on Twitter Academic API[3] to search drought-related tweets from 2012 to 2023. In particular, we consult with domain experts (i.e., researchers in civil and environmental engineering) and compile a list of relevant keywords (e.g., "drought", "rainfall", "wildfire", "famine", "groundwater", "wetland") to retrieve posts from Twitter. To protect user privacy and adhere to ethical standards, for each retrieved post, we only record the post ID, post text, publish timestamp, and geolocation (if available). In addition, we only consider posts written in English in our study. The summary of the

[3]https://developer.twitter.com/en/use-cases/do-research/academic-research

dataset will be discussed in the next section.

News Media Data

We collected drought-related news articles from various online news publishers. In particular, we adopt Google News[4], a popular news aggregation platform, to gather a comprehensive range of drought-related news articles. For each article, we crawl the title, article text, and its publicly accessible URL. Moreover, to ensure the quality and accuracy of the news article data, domain experts in our team manually verify the content and assign a categorical label about the article topic (e.g., "strategy", "status", "technology") for each article. Since the article verification and category annotation process requires domain expertise, we primarily focus on the news articles published after we launch the project, ensuring the domain experts have the most up-to-date knowledge to provide accurate verification and annotation. In addition, similar to the social media posts, we only consider news articles written in English in the news media data.

Meteorology Data

We consider the Drought Severity and Coverage Index (DSCI) as our primary meteorological metric for evaluating the extent and intensity of drought conditions in the studied areas. DSCI effectively compiles various indicators such as precipitation deficits, soil moisture, and meteorological conditions to assess drought severity (Smith et al. 2020). In particular, we collect DSCI statistics from the U.S. Drought Monitor, a comprehensive and authoritative source that tracks and analyzes drought patterns and severity across different regions of the United States. We retrieve the weekly DSCI for 51 regions in the U.S., including 50 U.S. states and Washington D.C., from January 2012 to December 2023.

Cleaning

We notice that the collected raw social and news media data contain irrelevant contents (e.g., off-topic posts, URLs) that need to be filtered out to ensure the dataset quality. To this end, we perform several data cleaning strategies to remove irrelevant social media posts and inaccurate content in web-crawled news articles. We elaborate on the details below.

Social Media Data We observe that the social media posts directly collected from Twitter (i.e., raw posts) often contain a non-trivial amount of irrelevant posts that are not pertinent to the natural drought in this study, mainly due to the ambiguity of the keywords in the search query. For example, the keyword "drought" also means periods of time without success or progress in sports, business, or creativity. A number of collected posts discussing "being in a scoring drought" or "a creative drought" are irrelevant to the context of the natural drought we are studying. Therefore, we develop a coarse topic model to cluster the social media posts into high-level topics and filter out posts in the off-topic clusters. In particular, we randomly sample 10% of the entire social media posts and train a BERTopic model (Grootendorst 2022), a state-of-the-art neural topic modeling approach, to identify major topics in the sampled social media posts. Among the 50 identified high-level topics, we manually verified and selected 14 topics that are relevant to the natural drought in this study. We apply the trained topic model to the entire set of social media posts and filter out irrelevant posts in off-topic clusters. Finally, we obtain a clean set with a total of 1,636,199 social media posts. A detailed statistical summary of the social media data is presented in the next section.

News Media Data We further clean the collected news articles to ensure their relevance and accuracy. Unlike the massive amount of social media posts, news articles are often written and edited by journalists, and the overall volume is much smaller. Therefore, we invite domain experts in our team to manually review and verify the content in each article. An article will be excluded if it contains irrelevant or factually incorrect/unverifiable information. We finally obtain 1,482 news articles from 523 news publishers. We discuss details of the news media data in the next section.

	Social Media Posts
Post Period	Jan. 2012 – Apr. 2023
Number of Raw Posts	3,562,605
Number of Relevant Posts	1,636,199
Keywords	drought, rainfall, wildfire, famine, groundwater, wetland

Table 2: Summary of Social Media Data

Data Summary

Statistical Summary

We summarize the key statistics of the social media, news media, and meteorology data in Table 2, Table 3, and Table 4, respectively. In particular, the social media data consists of 1,636,199 valid posts, 47,849 of which contain geolocations. Moreover, the news media data contains a total of 1,482 articles from 523 unique news publishers. The news articles are manually assigned into 29 categories. In addition, the meteorology data includes 31,977 weekly DSCI records in the U.S. from 2012 to 2023. The DSCI between 2012 and 2023 ranges from 0 to 476 with an average value of 79.89.

Content Summary

We present a summary of the textual content in the social and news media data. In particular, we plot the word clouds in Figure 3a and Figure 3b to characterize and visualize the most frequent words in social media posts and news articles, respectively. We observe that, in addition to the keywords used to retrieve the posts and articles, societal impact related terms (e.g., "people", "farmer", "food", "help") appear more frequently in social media posts. In contrast, news articles often contain event descriptive terms, such as "said",

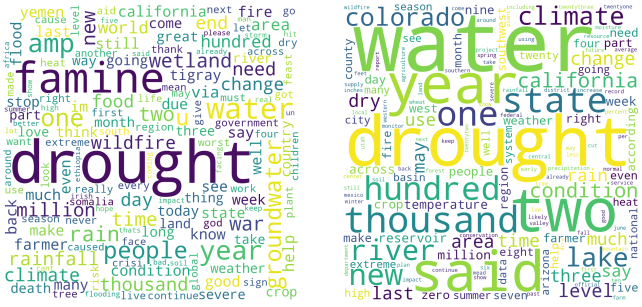
[4]https://news.google.com/

	News Articles
Collection Period	Jan. 2022 – Dec. 2023
Number of Articles	1,482
Number of News Publishers	523
News Categories	29
Average Text Length	8,474.3

Table 3: Summary of News Media Data

	DSCI Statistics
Collection Period	Jan. 2012 – Dec. 2023
Geographical Region	50 U.S. States, Washington D.C.
Measurement Frequency	Weekly
Number of Records	31,977
Mean DSCI	79.89
Range of DSCI	0 – 476

Table 4: Summary of Meteorology Data



(a) Social Media (b) News Media

Figure 3: Word Cloud

"condition", "river", and "lake", showing a more factual reporting style. Such an observation not only demonstrates the importance of social media posts in understanding the societal impact posed by drought but also highlights the complementary strengths of social media and news media in providing a comprehensive view of drought and its impact.

Geographical Distribution

We further visualize the geographical distribution of the geolocated social media posts and the meteorological DSCI records in Figure 4 and Figure 5, respectively. In particular, for the geolocated social media posts, we plot the normalized amount of posts in each U.S. state for each year from 2012 to 2023. For the meteorological DSCI records, we plot the average weekly DSCI of each U.S. state for each year from 2012 to 2023. We observe that the amount of geolocated social media posts is not necessarily correlated with the physical condition of drought severity denoted by DSCI. A possible reason is that social media data reflects human perception of drought and often depends on factors beyond just the physical drought conditions, such as population density and economic structure. For example, while the drought conditions in California in 2019 are less severe than in other southwestern states (e.g., Utah, Arizona), California is still the state that generated the most social media posts about drought that year due to its large population, agricultural industry, and water policy debates. Such an observation further demonstrates the necessity of integrating the insights from both physical measurement and human perception data to comprehensively understand the societal impact of drought.

Dataset Availability

We follow the FAIR Data Principles (FORCE11 2020) to release the SocialDrought dataset. The SocialDrought dataset is *findable* with a unique Digital Object Identifier (DOI): https://doi.org/10.5281/zenodo.10516342. The SocialDrought dataset is *accessible* through an open-access data repository Zenodo. The SocialDrought dataset is *interoperable* in that all the data are stored in Comma-Separated Value (CSV) files, a standard format for tabular data, and are readable with common data analytical tools (e.g., Python) The SocialDrought dataset is *re-usable* under a Creative Commons Attribution 4.0 International license, with an included README file and a Datasheet for the Dataset (Gebru et al. 2021) explaining the proper usage, the DOI to this paper (upon publication), and an open license allowing reuse and redistribution.

Preliminary Analysis

We perform a series of preliminary analyses to get an in-depth understanding of the collected SocialDrought dataset. The analyses include: 1) *Topic Analysis* that explores the major topics in the social and news media data and their dynamic changes, and 2) *Sentiment Analysis* that examines the sentiments of the social media posts and their relations with the physical severity of drought. In addition, we also develop an online analytical platform that is capable of collecting real-time social media and news media data based on customizable queries to gain timely insights into the societal impact of evolving drought conditions.

Topic Analysis

We first study the topics in social media posts and news articles to understand the major drought-related themes in online discourse. To extract topics in the social media and news media data, we train an unsupervised topic model for each dataset (i.e., social media, news media) using BERTopic (Grootendorst 2022), the state-of-the-art neural topic model architecture. First, we investigate the topic differences between social media and news media data. In particular, we show the top 5 topics and their semantic distance in terms of topic embeddings for the social media and news media data in Figure 6a and Figure 6b, respectively. We observe that, in addition to the shared topics (e.g., "tree", "forest") in

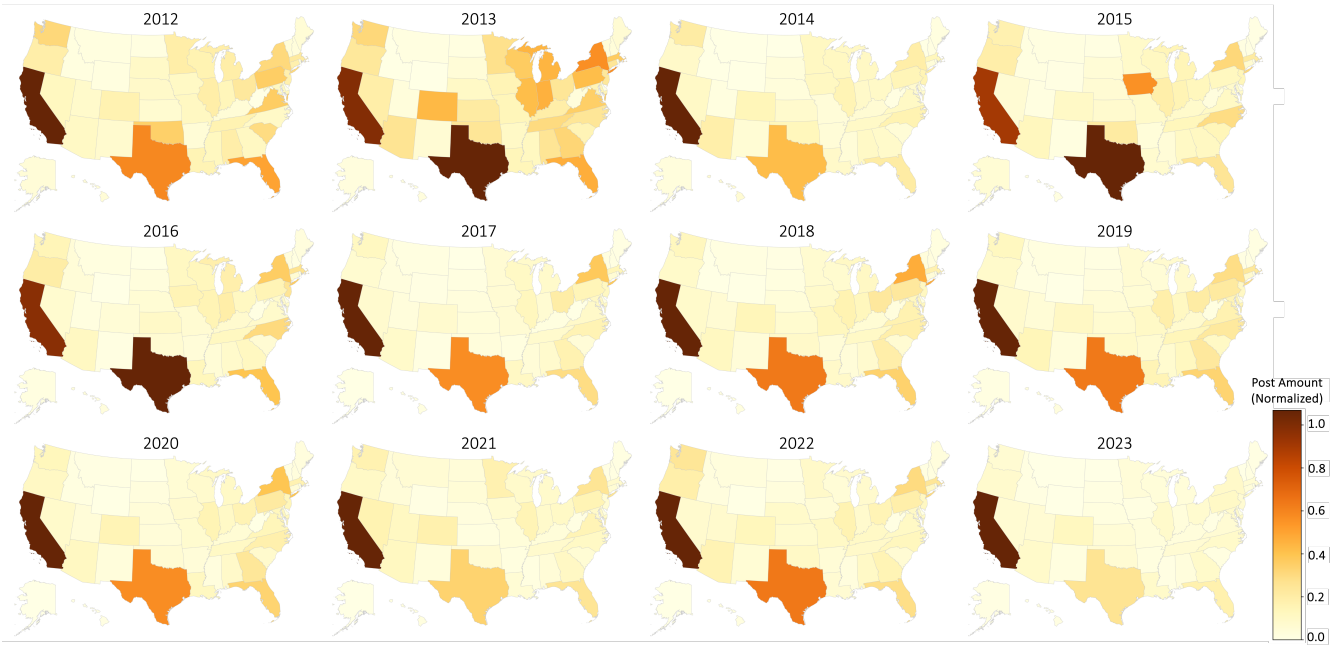


Figure 4: Geographical Distribution of Social Media Posts

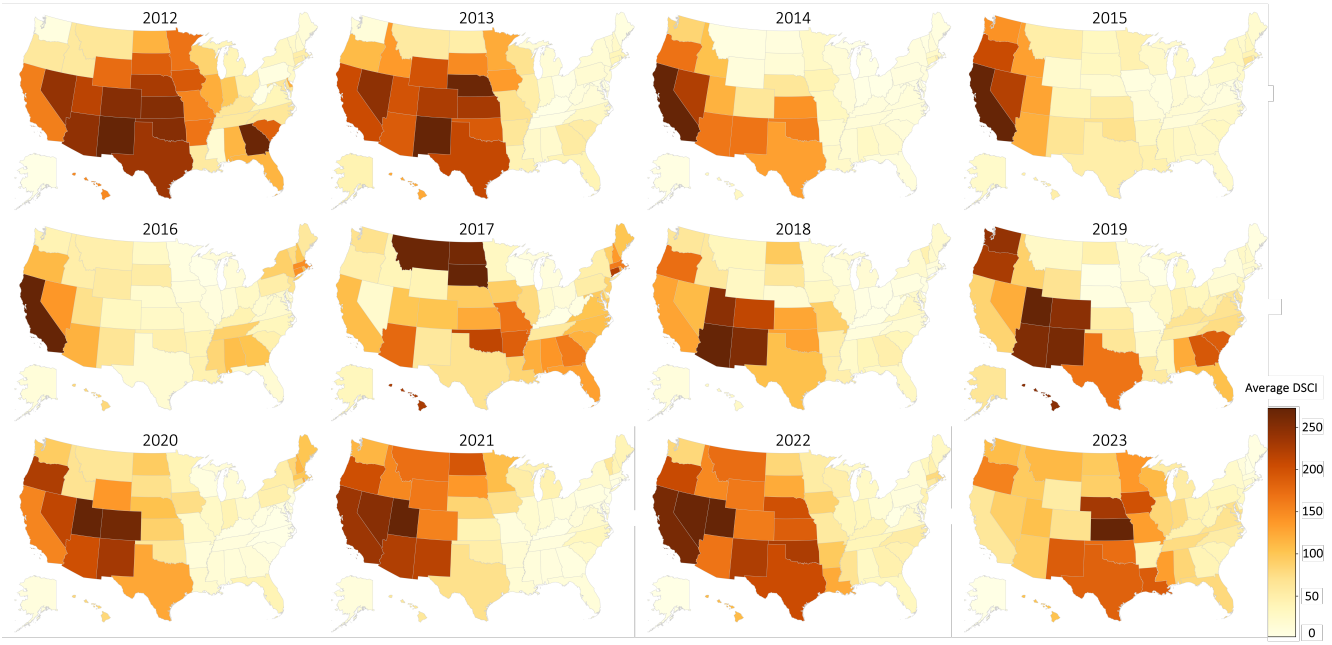
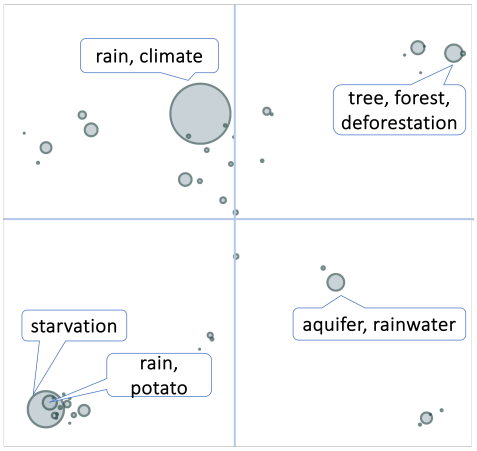


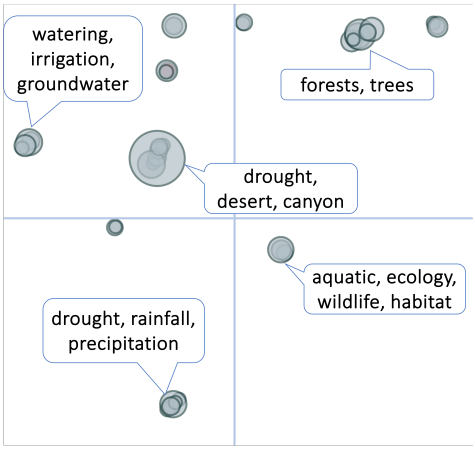
Figure 5: Geographical Distribution of DSCI

both social and new media data, the social media topics focus more on perceivable experiences or observations related to the drought, such as “rain/rainwater”, “climate”, “potato” while the news media topics tend to report on government actions/policies (e.g., “irrigation”, “watering”) and expert perspectives (e.g., “ecology”, “desert”) on the drought impacts. Such a difference highlights the advantages of social media data in capturing public observations and perceptions of drought impact.

Second, considering that social media data can better reflect human perceptions regarding the societal impact of drought than news media data, we further explore the temporal dynamics of topics in social media data to gain insights into how public attention and concerns evolve over time. Figure 7 shows the weekly frequency of social media posts over the past five years (from 2018 to 2023) for the top



(a) Social Media



(b) News Media

Figure 6: Intertopic Distance

10 topics identified by the topic model. We observe that the frequency of drought-related social media posts consistently increases over time. In addition, while the drought condition related topic of "rain and climate" remains the most concerning issue on social media, other societal impact related topics, such as "starvation", "bushfires and fires", "livestock, cattle, and cow", have shown rising trends in recent years, indicating the public's growing attention and concern regarding the societal impacts of the ongoing drought. The increases in public attention on these impact-related discussion topics highlight the value of social media data in sensing the societal impacts of drought and its related issues.

Sentiment Analysis

We further analyze the sentiments of social media posts with respect to their spatial and temporal distributions. In particular, we first train a sentiment classification model by fine-tuning the advanced RoBERTa model (Liu et al. 2019) on the TweetEval benchmark (Barbieri et al. 2020), a manually annotated Twitter dataset for sentiment analysis. The trained model is then applied to infer the sentiments of the geolocated social media posts in the SocialDrought dataset. We plot the average sentiment score of the posts in each U.S. state for each year between 2012 and 2023 in Figure 8. We observe that the dynamic variation of average sentiment often correlates with the changes in drought conditions in many states, especially for states with high population density and water-dependent agriculture/industries. For example, the average sentiment in Texas appears to be more negative during the drought years (e.g., 2012, 2022) as compared to the non-drought years (e.g., 2015, 2016). This indicates that social media posts can capture public sentiments reflecting drought's impacts on human society.

Online Analytical Platform

We develop an online analytical platform (Figure 2b) to facilitate the exploration and analysis of the real-time social and news media data, and the recent meteorological records to provide timely insights about drought and its impacts. In particular, the online analytical platform has the following functions. First, it shows the latest drought severity and coverage in the U.S. as the DSCI measured by the U.S. Drought Monitor. Second, it previews the social media and news media data collected in the SocialDrought datasets. Third, it contains a highlight page that shows the preliminary analysis results (e.g., topics and sentiments) of the SocialDrought dataset. Fourth, it features an insight page that is capable of retrieving real-time social media posts and news articles with customizable queries to support the investigation of emerging events and public opinions related to drought impact. Finally, the platform also allows users to share useful information and their findings to support more effective monitoring of drought societal impact. As the development of the online platform is ongoing, we will keep improving the functions of the platform and adding additional features to enhance its capabilities.

Experimental Settings

We preprocess all the text data in the preliminary analysis by removing the stopwords, URLs, mentions (i.e., @username), hashtags (i.e., #hashtag), punctuation, and special characters. We tokenize the text data using the standard BERT tokenizer (Song et al. 2020) and adopt the pre-trained RoBERTa embedding (Liu et al. 2019) with a vector length of 768 as the word representation. The default number of topics for the topic model is 50. The preliminary analysis of the SocialDrought dataset is conducted on 4 Nvidia A16 GPUs and the online analytical platform is hosted by the web-hosting service at the authors' institution.

Limitation and Discussion

In this study, we focus on the United States as our primary geographical area for our data collection and analysis. This is because drought has been a recurring issue in the United States, especially in the southwestern U.S. which has suffered from megadrought for decades (Medellin-Azuara et al. 2022). More specifically, prolonged droughts have posed significant societal impacts to human society in the U.S., ranging from reduced agricultural productivity and income

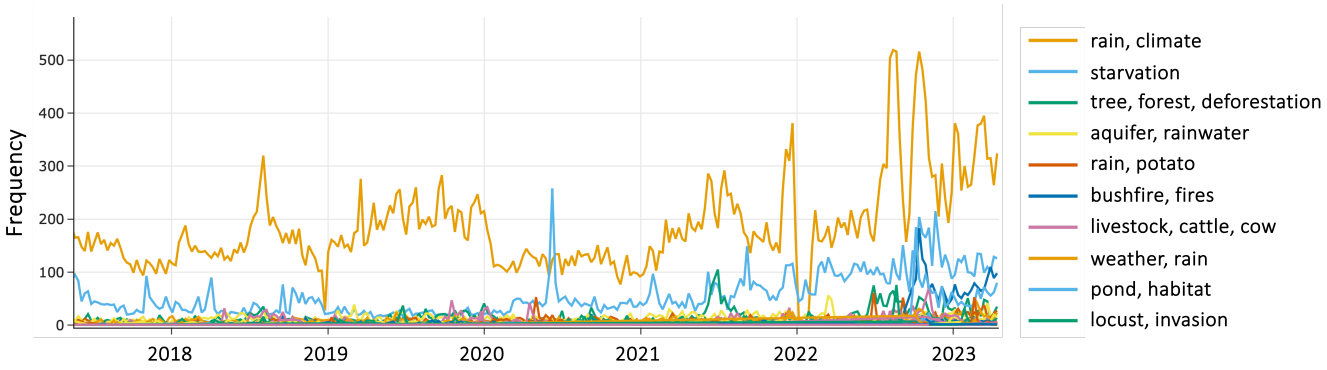


Figure 7: Topic Changes over Time

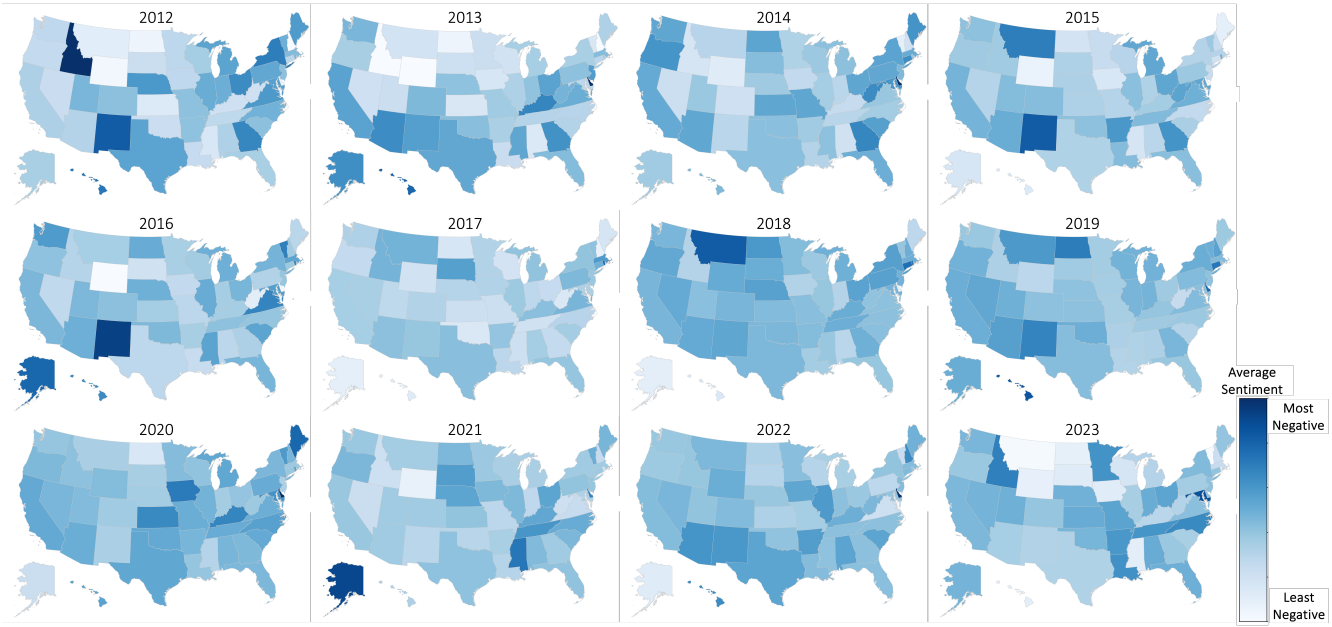


Figure 8: Geographical Variation of Social Media Sentiment

to public health concerns. Therefore, understanding the societal impact of drought in the U.S. can provide important insights for developing mitigation strategies, policies, and technologies to increase drought resilience in society. However, we believe the data collection strategy developed in this work can be deployed in other geographical regions (e.g., South America, Europe) to obtain useful information and localized insights about drought societal impacts. Moreover, this study primarily analyzes social and news media content in English which may result in information loss. In future work, we plan to incorporate machine translation techniques to gather a wealth of information for better understanding drought impact across different communities.

We note that the discontinuation of Twitter Academic API in Spring 2023 has significantly impacted social media data collection and analysis in recent studies. To mitigate such a challenge, we are developing several alternative strategies to continuously and responsibly collect social media data regarding the societal impact of drought. First, as discussed in the previous section, our online analytical platform has a dedicated function to collect input from online users regarding their experiences and observations on the societal impacts of drought. These crowdsourced user inputs will be further integrated into the SocialDrought dataset to provide more comprehensive views of drought impact. Second, we are exploring other social media platforms (e.g., Reddit, YouTube, TikTok) to supplement Twitter data and capture diverse drought discussions on social media. Third, we are adapting an API-free Twitter data scrapper that aims to directly collect publicly available tweets from Twitter's website while complying with Twitter's terms of service and ethical standards. We will keep exploring responsible and ethical methods to gather rich social media data on drought impacts as the landscape of drought evolves.

While this study aims to provide a comprehensive dataset capturing the societal impacts of drought, we acknowledge

several potential sources of bias. The SocialDrought dataset draws heavily from social media data, which has inherent biases in representing public perceptions. While social media data contain rich information on public perceptions from major populations, it may not fully represent the views of certain demographics, such as older adults who tend to be underrepresented on social media platforms. To mitigate such a bias, we plan to expand the social media platforms in our study by including data collected from other platforms (e.g., Reddit, YouTube) and the user inputs from our online analytical platform, to capture a greater diversity of ages, backgrounds, and perspectives. In addition, we also plan to collaborate with social science experts (e.g., sociologists) and survey organizations (e.g., Pew Research Center) to integrate findings from national surveys into our dataset, making our data sources more inclusive and representative.

Moreover, we notice that only a small portion (i.e., approximately 3%) of the social media posts contain geolocation information. The sparsity of geotagged social media posts may potentially limit the in-depth spatial analysis of public discussions and sentiment around the societal impacts of drought. To overcome such a limitation, a possible solution is to leverage natural language processing techniques (e.g., Serere, Resch, and Havas (2023), Fan, Wu, and Mostafavi (2020)) to infer location entities mentioned in the post text and approximate geolocation to enable richer geospatial analysis. Such a text-based strategy only accesses the publicly available post content and does not require geotags from the user end (e.g., user profile, device information), thereby ensuring user privacy and ethics around social media data analysis. However, such an approach may fall short of accurately capturing the true locations of all posts, especially for users who do not explicitly mention location-related terms. We plan to further investigate this problem and explore hybrid geolocation inference methods combining text-based location extraction and auxiliary information (e.g., anonymized user interactions) to achieve more robust location inference performance.

While we are committed to releasing the SocialDrought dataset, we acknowledge that there might be potential misuses of the dataset which may lead to negative outcomes of the work. For example, certain posts could be traced back to target or profile specific individuals or groups. To prevent the potential misuse of our dataset and the possible negative outcomes, we responsibly release the SocialDrought dataset in the study by only providing the tweet IDs for the social media data, the URLs for the news articles, and the public record of the DSCI data from the U.S. Drought Monitor. Such a strategy not only ensures the data retrieved from tweet IDs or the article URLs are up to date, but also safeguards the user profile and other personally identifiable information could not retrieved when a user removes a post or a news publisher removes an article. We also create and release a Datasheet for the Dataset (Gebru et al. 2021) with the dataset to clearly describe the proper usage and limitations of the dataset. In addition, we will also release the source code for the preliminary analysis upon the acceptance of the paper to ensure the reproducibility of findings.

Conclusion

This paper presents a novel SocialDrought dataset with an online analytical platform to facilitate research in understanding and analyzing the societal impacts of drought. The SocialDrought dataset consists of 1,636,199 social media posts, 1,482 expert-verified news articles, and 31,977 meteorological records from the U.S. Drought Monitor. By integrating social media, news media, and meteorological indices, this dataset, along with the online analytical platform, enables new opportunities to study the societal effects of drought at a large scale.

Broader Impact and Ethical Statement

The societal impact of drought has been a critical issue in many regions and affected communities worldwide. This work provides an important dataset for understanding the societal impacts of drought. While this work only focuses on the societal impact of drought in the U.S., we envision that the data collection scheme could be further adapted to study the societal impact of drought or other environmental issues (e.g., wildfires, flooding, deforestation) in other regions.

The SocialDrought dataset presented in this work is collected from publicly available sources, including social media platform (i.e., Twitter) for social media posts, online news websites for news articles, and the U.S. Drought Monitor for DSCI records. The research protocol was approved by the Institutional Review Board (IRB) at our institution. We note that the social media data may contain personally identifiable information of the social media users, leading to potential negative societal impact regarding user privacy. To ensure the ethics of the study and minimize such negative societal impacts, we only collected the post information (i.e., tweet ID, post content, and geolocation (if available)). To preserve user privacy, we did not obtain the personally identifiable information associated with each post (e.g., username, user profile). To comply with the ethical standards and terms of service, we will only release the tweet IDs for the social media data collected in this study.

Acknowledgments

We are grateful to Justin Xiao, Jimmy Miao, Ruozhen Yang, Candice Chen, Hang Qing He, Tianhao Chen, and Daniel Yao at the University of Illinois at Urbana-Champaign for helping with developing the online platform and collecting the drought-related data for the SocialDrought project.

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105032, IIS-2130263, CNS-2131622, CNS-2140999, and Ben Chie Yen Professorship fund for Professor Ximing Cai at the Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Barbieri, F.; Camacho-Collados, J.; Neves, L.; and Espinosa-Anke, L. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Becker, S.; and Sparks, P. 2020. "It never rains in California": Constructions of drought as a natural and social phenomenon. *Weather and Climate Extremes*, 29: 100257.

Dantas, J. C.; da Silva, R. M.; and Santos, C. A. G. 2020. Drought impacts, social organization, and public policies in northeastern Brazil: a case study of the upper Paraíba River basin. *Environmental monitoring and assessment*, 192: 1–21.

Edwards, B.; Gray, M.; and Hunter, B. 2019. The social and economic impacts of drought. *Australian Journal of Social Issues*, 54(1): 22–31.

Fan, C.; Wu, F.; and Mostafavi, A. 2020. A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access*, 8: 10478–10490.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2024-01-15.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Lai, C.; Zhong, R.; Wang, Z.; Wu, X.; Chen, X.; Wang, P.; and Lian, Y. 2019. Monitoring hydrological drought using long-term satellite-based precipitation data. *Science of the total environment*, 649: 1198–1208.

Lee, J.-W.; Hong, E.-M.; Jang, W.-J.; and Kim, S.-J. 2022. Assessment of socio-economic drought information using drought-related Internet news data. *International Journal of Disaster Risk Reduction*, 75: 102961.

Lester, L.; Flatau, P.; and Kyron, M. 2022. Understanding the Social Impacts of Drought. *Perth: The University of Western Australia*.

Lindersson, S.; Brandimarte, L.; Mård, J.; and Di Baldassarre, G. 2020. A review of freely accessible global datasets for the study of floods, droughts and their interactions with human societies. *Wiley Interdisciplinary Reviews: Water*, 7(3): e1424.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, K.; Wang, X.; de Jong, M.; and Flannigan, M. 2024. Drought triggers and sustains overnight fires in North America. *Nature*, 627(8003): 321–327.

MARIC, V. 2022. Analyzing drought impacts through social media and geophysical parameters. *Tesi di laurea Magistrale*.

Medellin-Azuara, J.; Escriva-Bou, A.; Rodriguez-Flores, J. M.; Cole, S. A.; Abatzoglou, J.; Viers, J. H.; Santos, N.; Summer, D. A.; Medina, C.; Arevalo, R.; et al. 2022. Economic Impacts of the 2020–22 Drought on California Agriculture.

Parsons, D. J.; Rey, D.; Tanguy, M.; and Holman, I. P. 2019. Regional variations in the link between drought indices and reported agricultural impacts of drought. *Agricultural systems*, 173: 119–129.

Salvador, C.; Nieto, R.; Linares, C.; Díaz, J.; and Gimeno, L. 2020. Effects of droughts on health: Diagnosis, repercussion, and adaptation in vulnerable regions under climate change. Challenges for future research. *Science of the Total Environment*, 703: 134912.

Serere, H. N.; Resch, B.; and Havas, C. R. 2023. Enhanced geocoding precision for location inference of tweet text using spaCy, Nominatim and Google Maps. A comparative analysis of the influence of data selection. *Plos one*, 18(3): e0282942.

Shang, L.; Kou, Z.; Zhang, Y.; Chen, J.; and Wang, D. 2022a. A privacy-aware distributed knowledge graph approach to qois-driven covid-19 misinformation detection. In *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*, 1–10. IEEE.

Shang, L.; Zhang, Y.; Ye, Q.; Wei, N.; and Wang, D. 2022b. Smartwatersens: a crowdsensing-based approach to groundwater contamination estimation. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, 48–55. IEEE.

Smith, K. H.; Tyre, A. J.; Tang, Z.; Hayes, M. J.; and Akyuz, F. A. 2020. Calibrating human attention as indicator monitoring# drought in the Twittersphere. *Bulletin of the American Meteorological Society*, 101(10): E1801–E1819.

Song, X.; Salcianu, A.; Song, Y.; Dopson, D.; and Zhou, D. 2020. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*.

Spinoni, J.; Barbosa, P.; De Jager, A.; McCormick, N.; Naumann, G.; Vogt, J. V.; Magni, D.; Masante, D.; and Mazzeschi, M. 2019. A new global database of meteorological drought events from 1951 to 2016. *Journal of Hydrology: Regional Studies*, 22: 100593.

Williams, A. P.; Cook, B. I.; and Smerdon, J. E. 2022. Rapid intensification of the emerging southwestern North American megadrought in 2020–2021. *Nature Climate Change*, 12(3): 232–234.

Zhang, Y.; Shang, L.; Zong, R.; Wang, Z.; Kou, Z.; and Wang, D. 2021a. StreamCollab: A streaming crowd-AI collaborative system to smart urban infrastructure monitoring in social sensing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 179–190.

Zhang, Y.; Zong, R.; Shang, L.; Kou, Z.; and Wang, D. 2021b. A deep contrastive learning approach to extremely-sparse disaster damage assessment in social sensing. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 151–158.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, please see the Limitation and Discussion section.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, please see the Abstract and Introduction sections.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, please see the Data Collection section.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, please see the Limitation and Discussion section.
   (e) Did you describe the limitations of your work? Yes, please see the Limitation and Discussion section.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, please see the Broader Impact and Ethical Statement section.
   (g) Did you discuss any potential misuse of your work? Yes, please see the Limitation and Discussion section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, please see the Limitation and Discussion section.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we have carefully reviewed the guidelines and make sure our work conforms to them.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, according to the submission guideline, we provided a small sample of the dataset as the supplementary material. The full dataset is published on Zenodo at: https://doi.org/10.5281/zenodo.10516342.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, please see the Data Collection section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, please see the Preliminary Analysis section.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, please see the Data Collection and Preliminary Analysis sections.
   (b) Did you mention the license of the assets? Yes, please see the Data Summary section.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, please see the Data Summary section.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, please see the Data Collection section.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, please see the Data Collection and Ethical Statement section.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, please see the Limitation and Discussion section.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, we created a Datasheet for the Dataset and will be released with the dataset.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA