

An Analysis of the Jackson Mississippi Water Crisis using Social and News Media


Rishabh Shrestha¹, Ting Xiao², and Dawn E. Wilkins^{1,2}

¹ Department of Computer and Information Science, University of Mississippi

² Institute for Data Science, University of Mississippi

Author Note

Ting Xiao  <https://orcid.org/0000-0002-7502-4395>

Dawn E. Wilkins  <https://orcid.org/0000-0001-6047-7327>

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Dawn E. Wilkins,
Department of Computer and Information Science, 203 Weir Hall, University, MS 38677.

Email: dwilkins@olemiss.edu

Abstract

In this paper, we present an analysis of social media posts and mass media communications during the 2021 Jackson Water Crisis. The focus of this research is to analyze tweets and news media data to identify the public discussions and their sentiments towards the Jackson Water Crisis. We performed data analysis to identify what people were talking about on Twitter regarding the Jackson Water Crisis and how they were feeling towards the issues raised from the water crisis. Topic modeling helped to identify different major topics and tweets were categorized into four major topics. Sentiment analysis helped to find the positivity and negativity of the tweets and news articles.

Keywords: Jackson Water Crisis, social media, news media, tweets, sentiment analysis, topic modeling, machine learning

Background

Recently there has been a transition in how citizens acquire news and information about important events. Traditionally, print and broadcast media played a major role in determining what events to cover and how to shape stories for the audience. In recent years, social media has changed the way most people consume information about current events (Hermida, 2010). Social media has also allowed citizens to join in the conversation and contribute to the discussion in real-time. Today, when there is a crisis or significant event, citizens take to social media, especially Twitter, to post facts, opinions, concerns, suggestions and, sometimes, misinformation (Vicario et al., 2016). During the Jackson Water Crisis, many people used social media to ask for help, prayers, and donations. They used a range of hashtags, including #JacksonWaterCrisis, #JxnNeedsWater, #JxnWater, #EnvironmentalJustice, #infrastructure, and others. This data is a largely untapped resource of valuable information for community leaders and first responders.

We analyzed both social media posts and online, but traditional mass media related to the Jackson Water Crisis. Data were collected and have been statistically summarized to align with the timeline of the crisis. Sentiment analysis and topic modeling of the social media posts also provide a high-level understanding of the impact of the crisis (Wang & Taylor, 2018; Xiong et al., 2020) and the concerns of the citizens.

Research Methods

Data Collection and Preprocessing for Twitter

We used Twitter as our primary data source of public opinion. Tweets related to the Jackson Water Crisis from February 10th, 2021 to May 20th, 2021 were accessed and collected through the snsrape library, a web scraping tool for social networking services (SNS), in Python. We used Python and R as our programming languages. We scraped data using search criteria such as keyword, start, and end date. In addition to that, we used geocode and radius in some queries to collect tweets from the Jackson, MS area.

To collect tweets that are most relevant to the Jackson Water Crisis, we used hashtags and keywords such as ‘#Jacksonwatercrisis’, ‘#Jacksonneedswater’, ‘#Jacksonwater’, ‘Jackson MS water problem’, ‘Jackson water crisis’, ‘Jackson water pipe burst’, ‘Jackson water problem’, ‘Jackson water outage’, and ‘Mississippi water crisis’. Some of the data were collected using general keywords such as ‘water pipe’, ‘frozen pipe’, ‘water outage’, ‘water crisis’, ‘water problem’, and ‘water shortage’, along with the date and the geocode (32.2998 latitude and -90.1848 longitude with radius of 50 km), to collect tweets made from the Jackson area.

For each tweet that matched the search criteria, we collected the tweet text, tweet ID, date and time tweet was posted, user name for the person who tweeted, user location, likes count, replies count, quotes count, and retweet count. Moreover, we collected the related tweets for the

water crisis in Jackson since January 1st, 2010 to explore whether there were similar problems in the past.

The tweets were investigated and duplicates were identified and removed for further analysis. For sentiment analysis, we used raw tweet texts. After that, the tweet texts were cleaned by removing special characters, symbols, numbers, and hyperlinks. The tweet texts were further tokenized and a list of words from the tweet was created for each tweet. Stop and noise words, which do not add much value to our analysis, were also removed. The preprocessed tweets were used for Topic Modeling and Machine Learning.

Data Collection and Preprocessing for News Media sources

We collected data from news media sources from February 10th, 2021 to June 20th, 2021 via Google News Application Programming Interface (API) in Python. We used search terms ‘Jackson Mississippi water crisis’ and ‘Jackson Mississippi water outage’, ‘Nick Judin Jackson water crisis’. We also accessed some news articles via searches in Google and other online links. We collected the date published, media source, the title of the article, and article content for each article retrieved.

The news data were cleaned and preprocessed before performing analysis on their textual content. The articles were cleaned by removing special characters, symbols, numbers, and hyperlinks before doing further analysis such as Topic Modeling.

Exploratory Data Analysis

The data collected were visualized in various ways. We examined the tweets and news articles over time and by location. We explored the most frequent words in the tweets and news articles. VADER (Valence Aware Dictionary and Sentiment Reasoner) was used for sentiment analysis since it works very well with short texts and texts from social media. VADER also

provides a compound sentiment score of texts ranging from -1 to +1 which helps us identify whether a given text has positive, negative or neutral sentiment (Hutto & Gilbert., 2014).

Similarly, we again used VADER for consistency to perform sentiment analysis on news articles.

Topic Modeling

Topic modeling is a natural language processing technique and an unsupervised machine learning method that scans a set of documents (or tweets), detects common words and phrase patterns, and then automatically clusters similar words and phrases together that best characterize the set of documents. It is a type of statistical model for discovering the abstract “topics” that occur in a collection of documents. To classify the tweets collected, we applied Latent Dirichlet Allocation (LDA) as our topic modeling method which is a very effective and commonly used method for topic modeling (Jelodar et al., 2019). We used preprocessed tweets as described above to perform topic modeling on them. The text data were tokenized, lemmatized, and stop words were removed before implementing LDA (Azad, 2020).

The LDA model returned the topics in clusters of words along with their corresponding weights. The number of topics to be returned was manually set to obtain the optimal number of topics, which we set to four.

After getting the topics as a cluster of words, we chose one single representative label for each topic which effectively describes the topic. For that, we used the following rules: (Xiong et al., 2020).

- a. Redundant terms were removed from other topics
- b. Stopwords were removed and keywords explicitly used to collect data were filtered out
- c. Words that could not be found in VADER word collections were excluded
- d. Lemmatized words were used

- e. The synonyms were compared with their weights and the highest was chosen
- f. Verbs that did not add much meaning to the topic were excluded
- g. The top three weighted words were selected after applying all the above rules
- h. From the top 3 words, the one with the highest weight and that could generalize the entire topic was chosen

For example, below is one of the topics returned by the LDA model.

$0.018 \times \text{"week"} + 0.014 \times \text{"infrastructure"} + 0.014 \times \text{"mayor"} + 0.013 \times \text{"help"} + 0.012 \times \text{"break"} +$
 $0.010 \times \text{"month"} + 0.009 \times \text{"racism"} + 0.008 \times \text{"pandemic"} + 0.008 \times \text{"people"} + 0.008 \times \text{"today"}$

After applying the above rules(a-h), we are left with the following words: infrastructure, break, and pandemic. From this we chose “infrastructure” as our one-word topic for this cluster(Rule h).

Labelling of Tweets

We used the four topics identified from topic modeling of the tweets to hand label each of the tweets. In addition, some tweets that didn't fit into one of the four topics were labelled None (N) and removed from subsequent processing. Many tweets could have been placed in multiple topic groups, but one primary topic was selected to reflect the primary motivation for the tweet.

Machine Learning

Once the tweets were labeled with one of the chosen categories, it is possible to use machine learning to predict labels for unseen tweets. To extract features from text documents (tweets), we used the Countvectorizer from scikit-learn library (Pedregosa et al., 2011), a feature extraction technique for text classification. It is used to transform texts into a vector on the basis of frequency of words in the text. CountVectorizer creates a matrix in which each unique word is

represented by a column of the matrix and each text sample(tweet) from the document is a row in the matrix. The value of each cell in a row is the count of the word in that row or text sample (Verma, 2020).

After implementing the CountVectorizer, we used columns in vector as features and three different machine learning models were built (using the sci-kit learn library in Python) – k-nearest neighbor, a decision tree model and a random forest classifier (Boehmke et al., 2020).

Results

Exploratory Data Analysis with Tweets

A total of 4005 tweets were collected. Figure 1 shows the number of tweets collected each day during the acute period of the water crisis. The most common “mention” in tweets was Governor Tate Reeves (@tatereeves), with 52 mentions.

Figure 1

Number of Related Tweets from February 10 to April 1, 2021

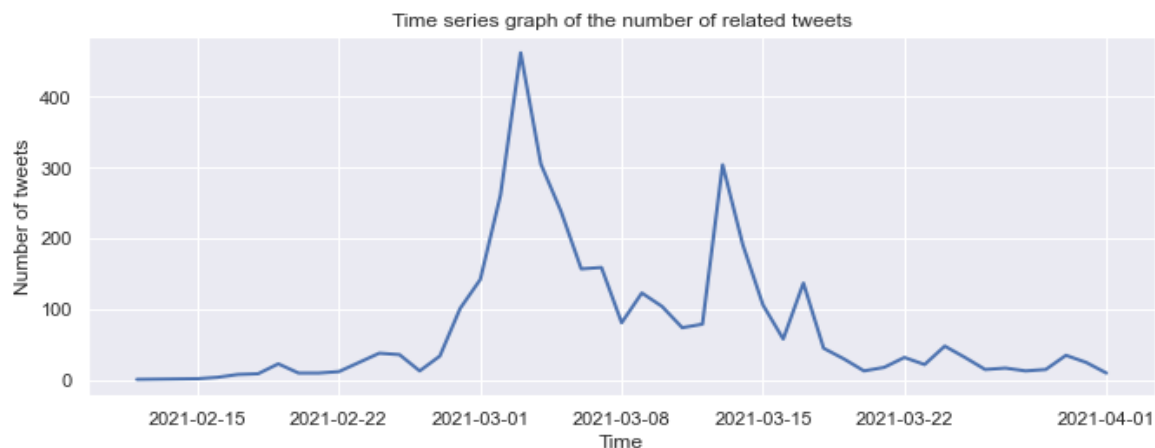
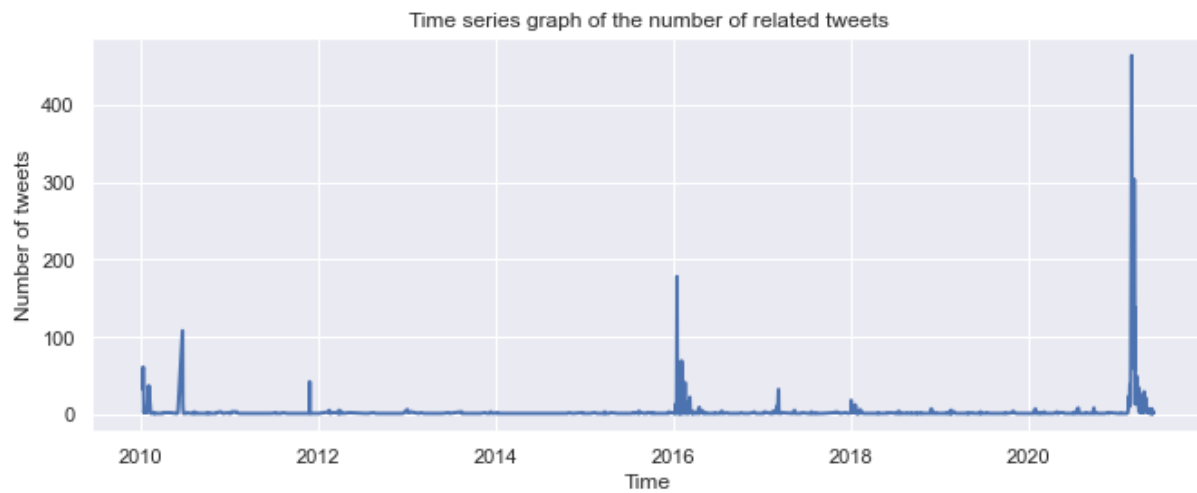


Figure 2 shows the number of tweets since January 1st, 2010 that mentioned the Jackson water crisis or problem. This clearly shows there have been a number of issues in the past (most notably in 2010 and 2016), but none were as bad as that in 2021.

Figure 2

Time Series Graph of the Number of Related Tweets from 2010 to 2020



While the vast majority of the tweets were from Mississippi, there were tweets from all over the country. Of the more than 4000 tweets collected, 75% had specified location information. Some tweets list city and state, others just state, or country, or Earth. Table 1 lists the top 8 cities where the most tweets about the water crisis were made along with their respective tweet counts. Of around 682 total tweets in Mississippi, around 399 were from Jackson, MS.

Table 1

Tweet counts by locations

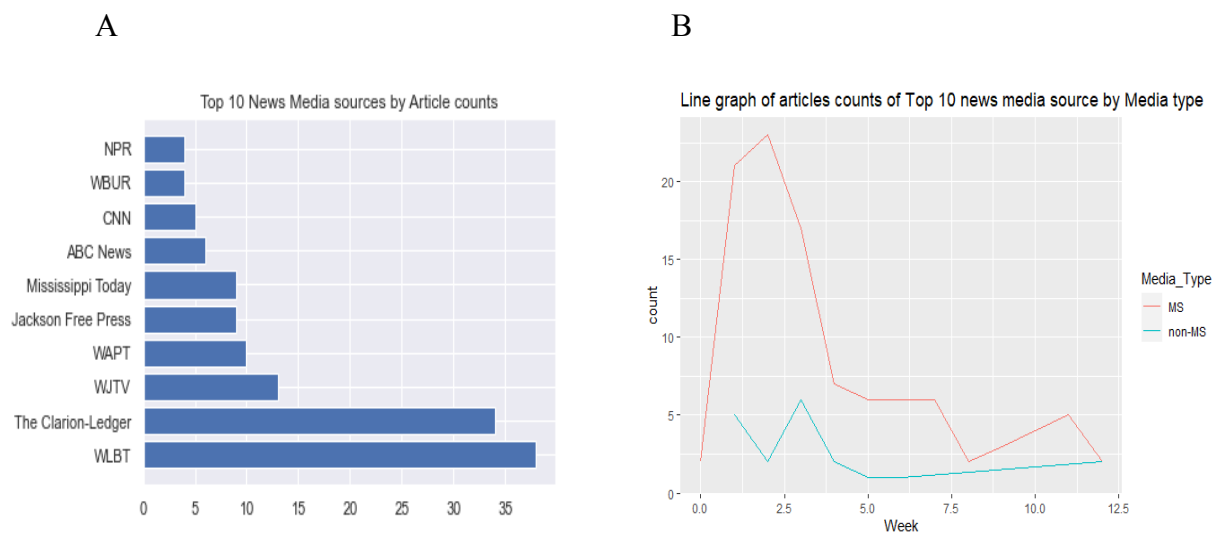
Location	Tweet count	Location	Tweet count
Jackson, MS	399	Chicago, IL	39
Los Angeles, CA	132	Hattiesburg, MS	32
New York, NY	104	Atlanta, GA	31
Washington DC	89	Biloxi, MS	21

Exploratory Data Analysis with News Articles

In total, the corpus of news media sources collected includes 224 articles from 78 different sources. Figure 3A shows the number of news articles about the Jackson Water Crisis by the top 10 news media sources on the basis of article counts in our collection. Figure 3B shows the counts over the weeks by primary location of media sources (within Mississippi, and outside of Mississippi). The majority of articles were published by Mississippi-based new sources, especially in the early weeks of the crisis.

Figure 3

Media Sources with Top 10 Highest Article Counts

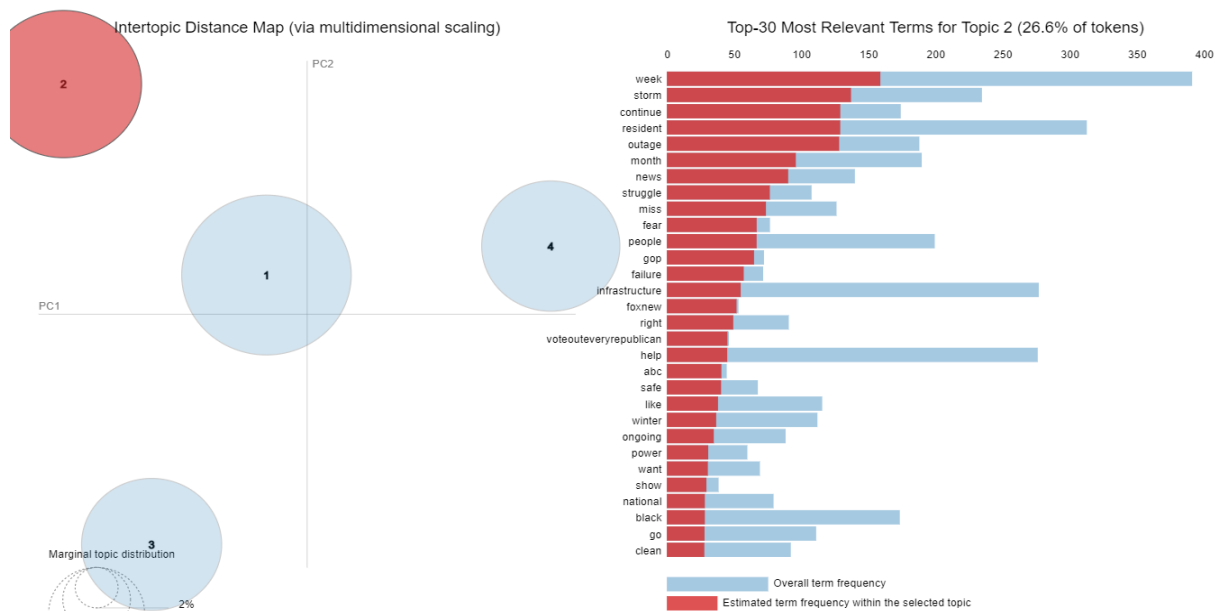


In Figure 4A, after removing some common words, such as Jackson, Mississippi, water, etc., a word cloud was produced using the text from all tweets. This gives an idea of the most common terms, where the larger the word the more occurrences found in the tweets. In Figure 4B, the word cloud produced from the news articles is similar to the tweet word cloud, except for a strong focus on “resident”, “system”, and “people” instead of “problem”, “city”, and “MS”. In

The returned topics were visualized using pyLDAvis which returns an Intertopic Distance Map and Top 30 most relevant terms from each topic as we can see in Figure 5. The intertopic distance map shows topics in the form of bubbles. The larger the bubble, the higher percentage of tweets in the corpus that are related to that topic. Each bubble represents a topic and the distance between bubbles represents how different topics are from each other.

Figure 5

Intertopic Distance Map of Topics



Using these four topics, we manually labelled each tweet and gave each one of the four topic labels. Some tweets were not related to any of the topics or were duplicates. These tweets were marked N (None) and removed from future analysis. A number of tweets mentioned multiple topics, but we used the label that best matched the main point of the person posting the tweet. Representative sample tweets from each topic are shown in Table 3.

Table 3*Number of Tweets and Sample Tweets in Different Topics*

Topic	Number of tweets	Sample tweets
Infrastructure	309	<p>“Clean safe water will be a thing of the past if we don't fix our infrastructure now.”</p> <p>“Sadly, this is a constant problem in Jackson. I lived in the city for 5 years and we were almost always under a boil water notice. Then in swoops a natural disaster that crushes the already broken infrastructure”</p>
Legislative / Political	373	<p>“The Mississippi Legislature’s effort to assist in the Jackson water crisis has been pared down to a single bill on water-payment flexibility and \$2 million from the capital expense fund.”</p> <p>“Shame, the governor @tatereeves is more concerned about banning trans youth from sports rather than focusing on actual issues...such as Jackson's water crisis!”</p>
Struggle	717	<p>“Mississippi residents are having trouble finding bottled water for sale. Distribution sites are set up now.”</p> <p>“AN ONGOING STRUGGLE: Mississippi's largest city is still struggling with water problems more than two weeks after winter storms and freezing weather ravaged the system in Jackson.”</p>
Racism	313	<p>“Black sections of Jackson, Mississippi, are nearing two weeks without water. White areas of town have no problems.”</p> <p>“The water crisis in Mississippi isn't shocking. It's a continuous cycle of environmental racism. #ADOS”</p>

Notes about the tweets:

- There were many tweets about articles written by Nick Judin of the Mississippi Free Press. This was the motivation to do an additional search of the news articles to ensure Mr. Judin’s articles were included in the corpus.
- Several tweets referenced “Nina Simone was right”, in reference to the civil rights songs by the singer, likely “Mississippi Goddam”.

- A few tweets used hashtag #ADOS or #MississippiADOS, referring to the American Descendants of Slavery movement. Most of these posts were categorized under the Racism topic.
- A number of tweets focus on Governor Reeves actions on transgender athletes or lifting mask mandate in the middle of the water crisis.
- Many tweets talked about Texas and their politicians, who were experiencing water issues based on the same winter storm that caused the problems in Jackson.
- Another commonly tweeted article was “Jackson, Mississippi has a water crisis because our state legislature has a race problem”, written by Donna Ladd (2021), founding editor of the Jackson Free press and the non-profit the Mississippi Free Press.

Machine Learning

Since we had hand-labeled the tweets using the four topics – Infrastructure, Legislative, Struggle and Racism, we wanted to see if we could predict the category of unseen tweets using machine learning. Tweets marked None were removed. The remaining 1712 cleaned and vectorized tweets included 717 (41.9%) categorized as Struggle, 373 (21.8%) that were marked Legislative, 313 (18.3%) labeled as Racism, and 309 (18%) listed as Infrastructure. For evaluation of the models, 80% of the tweets were used to train several different models and then each was tested on the remaining, unseen 20% of the tweets. Each model built was used to predict which of the four topics each unseen tweet best matched. The predictions were compared to their assigned labels to determine the accuracy of the model.

Three different machine learning models were built – k-nearest neighbor, a decision tree model and a random forest classifier. The k-nearest neighbor model had accuracy of about 53.4%, the decision tree model had accuracy of about 69.4%, and the random forest classifier

achieved about 74.9% accuracy. Given the overlapping nature of the topics, these are reasonable results. Being able to predict the categories would be useful in a real-time monitoring system to organize the crowd-sourced discussion for city managers to aid in decision making.

Topic Modeling of News Data

Topic modeling was also attempted using the news data. In Table 4, the four topics extracted from news data and their respective keywords are presented. The topics are not as clear as in the tweets. Topic 3 is closest to Struggle. Topics 2 and 4 have parts of Infrastructure and Legislative / Political. Topic 1 does not seem to match any of the topics from the tweets. Moreover, there is no significant discussion of Racism in the news articles, unlike the situation in the tweets. Given the more fact-based nature of the news articles relative to the more emotional tweets, this was expected.

Table 4

Topics in News Articles

Topic	Keywords
1	system residents williams pressure st road crisis still distribution mayor
2	state million system infrastructure crisis help mayor federal would need
3	power people state texas winter without residents weather still storm
4	state residents lumumba people infrastructure mayor system crisis also without

Sentiment Analysis

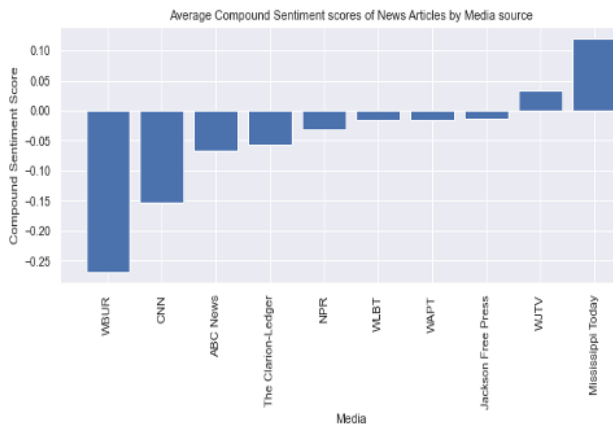
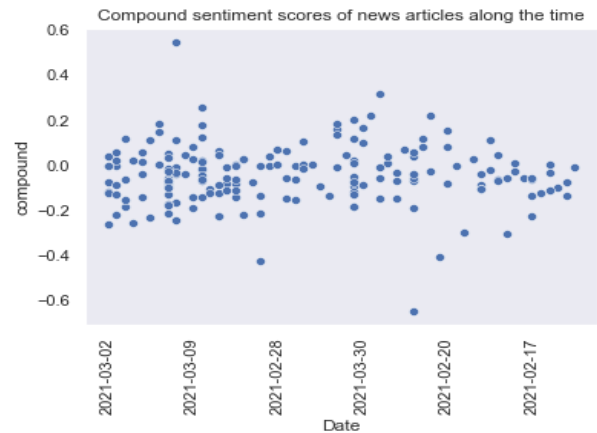
News media articles, which are generally written by journalism professionals, typically focus on facts and conveying accurate information. News articles typically display little sentiment (either positive or negative). On the other hand, tweets are posted by individuals and they have a much broader range of sentiment. This expectation was confirmed by the Jackson water crisis data we collected. Using VADER, both the tweets and the news articles were evaluated for sentiment. The measure used is the compound sentiment score -- the sum of the

positive, negative and neutral sentiment scores, which is then scaled to be between +1 (positive) and -1 (negative). Overall, using VADER, the news articles scored an average of -0.0426 and the average tweet score was significantly more negative, -0.4974. The average sentiment scores of the tweets by category were also all negative: Struggle -0.3317, Racism -0.4688, Legislative -0.5200, and Infrastructure -0.52235.

Figure 6 shows the compound sentiment of the news articles. Figure 6A shows the compound sentiment score by the news outlet, and Figure 6B shows the compound sentiment scores of each news article over time. There are a few outliers, but the majority of the articles have sentiment scores between +0.2 and -0.3. There was no significant change in compound sentiment over time. Interestingly, Mississippi Today was slightly positive, while WBUR (Boston's NPR station) and CNN were the most negative with respect to compound sentiment.

Popular n-grams

Using the cleaned tweets, with duplicates removed as much as possible, we collected the top 10 most frequent words from each of the four topics identified in the topic modeling step. Then we found the most frequent sets of two word phrases (bigrams) and three word phrases (trigrams) using those terms. Some of the most common bigrams were "crisis continues", "two weeks", "month without", "next outage", "struggling residents", "residents fear" and "winter storm". The common trigrams include "struggling residents fear", "crisis state legislature", "long standing problems", "crisis amid pandemic", and "month long crisis". These phrases succinctly summarize many of the tweets.

Figure 6*Compound Sentiment Scores of News Articles***A****B**

Discussion

Tweets are extremely valuable to monitor public opinion. Other social media platforms could be included for broader reach. We identified four main topics from the tweets – Struggle (41.9%), Legislative (21.8%), Racism (18.3%), and Infrastructure (18%). The number of tweets categorized about the Struggle was over twice as many as each of the other three topics, which all had similar numbers of tweets.

Strengths

- Analysis of social media can be done in real time, allowing decision makers and first responders to provide aid and assistance
- Location data is inconsistent, but if used with city or sub-city markers, could be extremely valuable in crisis tracking
- Identification of topics that are discussed and tracking them over time is possible
- Sentiment on the topics is quantifiable and can be passively determined

Limitations

- Cleaning the tweets and removing near duplicates and retweets is difficult. Sometimes the same tweet is posted multiple times with different mentions.
- Overlapping topics could be modeled better than attempted here
- News articles are often published with only minor changes from the original source and are not as valuable as social media to track a crisis in real time
- There was not enough data in this case study to see the spread of the story clearly, likely due to the relatively local nature of the story

Future Work

The work in this paper can not only inform policy decisions but more importantly lead to a city dashboard that can aid communities in real-time when a crisis occurs. A city dashboard could incorporate data from different sources. Typical data sources include local government data, public survey data, local service data, environment data, crowd-sourced data, and social or news media data (Kitchin et al., 2016). By building a city dashboard, a city can 1) keep citizens informed during a crisis, 2) provide real-time data for city managers and decision-makers, and 3) create a historical record to allow communities to be more resilient in managing future disasters better. Good examples include the dashboard for London, the United Kingdom (<https://citydashboard.org/london/>), and that for Dublin, the Republic of Ireland (<https://www.dublindashboard.ie/>). Moreover, features like crisis prediction and predictive alerts are able to be added to the dashboard. In commercial crisis dashboards like the NewsWhip dashboard, crisis prediction models were built to guide decisions daily (Quigley, 2020).

References

- Azad, A. (2020, July 12). *Twitter topic modeling*. Towards Data Science.
<https://towardsdatascience.com/twitter-topic-modeling-e0e3315b12e2>
- Boehmke, B., & Greenwell, B. (2019). *Hands-On Machine Learning with R (1st ed.)*. Chapman and Hall/CRC. <https://doi.org/10.1201/9780367816377>
- Hermida, A. (2010). TWITTERING THE NEWS: The emergence of ambient journalism. *Journalism Practice*, 4(3), 297-308. <https://doi.org/10.1080/17512781003640703>
- Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216-225.
<https://ojs.aaai.org/index.php/ICWSM/article/view/14550>
- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78(11), 15169-15211. <https://doi.org/10.1007/s11042-018-6894-4>
- Ladd, D. (2021, March 6). Jackson, Mississippi has a water crisis because our state legislature has a race problem. *NBC News*.
<https://www.nbcnews.com/think/opinion/jackson-mississippi-has-water-crisis-because-our-state-legislature-has-ncna1259819>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825-2830.
<https://dl.acm.org/doi/10.5555/1953048.2078195>

Quigley, P. (2020, October 21). Introducing crisis dashboard. *NewsWhip*.

<https://www.newswhip.com/2020/10/crisis-dashboard-release/>

Verma, K. (2020, July 17). Using CountVectorizer to extracting features from text.

GeeksforGeeks.

<https://www.geeksforgeeks.org/using-countvectorizer-to-extracting-features-from-text/>

Vicario, M. D., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., &

Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences - PNAS*, 113(3), 554-559.

<https://doi.org/10.1073/pnas.1517441113>

Wang, Y., & Taylor, J. E. (2018). Coupling sentiment and human mobility in natural disasters: A

twitter-based study of the 2014 south napa earthquake. *Natural Hazards (Dordrecht)*, 92(2), 907-925. <https://doi.org/10.1007/s11069-018-3231-1>

Xiong, J., Hsuen, Y., & Naslund, J. A. (2020). Digital surveillance for monitoring environmental health threats: A case study capturing public opinion from twitter about the 2019 Chennai water crisis. *International Journal of Environmental Research and Public Health*, 17(14),

5077. <https://doi.org/10.3390/ijerph17145077>