Uncertainties in Geolocating Social Media Data for Disaster Research

Debayan Mandal [a,*], Lei Zou [a], Heng Cai [a], Binbin Lin [a], Bing Zhou [a], Joynal Abedin [a], Mingzheng Yang [a]

[a] Department of Geography, Texas A&M University
* rohan_debayan@tamu.edu

Keywords: social sensing, uncertainty, geolocation, social media, disaster resilience

Introduction

Social sensing refers to the use of geospatial big data, e.g., social media, mobile phone, taxi data, and smart card data, to reveal socioeconomic characteristics (Liu et al. 2016). It supplements traditional remote sensing and survey data by providing more insight into human behaviors, spatial interactions, and place semantics of the society at an unprecedented scale in near real-time. Globally, social media usage coupled with Global navigation satellite system (GNSS)-enabled portable devices has become an indispensable part of daily life, which turns every user into a sensor capturing a direct snapshot of human activities at various places. In recent times, we have seen a recent rise in using social media data for tackling different societal studies, e.g., disaster resilience (Zou et al. 2018; Zou et al. 2019; Wang et al. 2015), Covid-19-related studies (Lin et al. 2022; Cinelli et al. 2020; Tsao et al. 2021; Wong et al. 2021), political sway, and religious and economic trends (Graham et al. 2014), etc.

Twitter posts can be geolocated through three attributes of the original Twitter data, (a) geotagged points or places reported by the device GNSS (referred to as tweet from), (b) locations mentioned in tweet contents (tweet about), and (c) user profile addresses (user from), but each method has limitations. Geotagged locations are considered ground truth representing user locations, but geotagged tweets form a small percentage of the data procured. As a result, spatial analysis relying on geotagged tweets may overlook most of the information.

In the absence of geotagged data, most researchers turn to geo-enriched data to get a reasonable sample size. Geo-enrichment (Dwoskin, 2014) of the Twitter data usually consists of two approaches. If any address is mentioned in the tweet (tweet about), the address is used to represent the user’s location. Alternatively, the user from location in the user profile can be considered. However, not every user update profile information promptly. On the other hand, the address mentioned in a tweet does not necessarily reflect the user’s current location.

The purpose of this study is to quantify the uncertainty of geolocated tweets in social media analysis for disaster research when ‘tweet about’ and ‘user from’ information is used. We collected Twitter data during the 2017 Hurricane Harvey, one of the deadliest hurricanes in U.S. history, causing billions of dollars of economic loss and intensive discussions on Twitter. This study has two specific objectives: (1) to develop a framework for visualizing accuracies of representing user locations by the ‘user from’ and ‘tweet about’ locations in different disaster phases at multiple spatial scales; (2) to showcase the implications of geolocating uncertainties in social media analysis for disaster research and management.

Method

Figure 1 presents the detailed workflow of this investigation. The first step was initial data collection. A total of forty-seven million Harvey-related tweets were retrieved by a list of pre-defined keywords [*harvey, hurricane, storm, flood, houston, txtf* (Texas Task Force), *coast guard, uscg* (U.S. Coast Guard), *houstonpolice, cajun navy, fema* (Federal Emergency Management Agency), *rescue*] from the Twitter Company. We filtered the tweets having geotags, resulting in a total of 588,401 tweets. The geotags were used as the ground truth for locations. The next step was geolocating collected tweets through 'tweet about' and 'user from' approaches. The user location information was extracted from the tweet details while for parsing the tweet content the python package "locationtagger" (Soni, 2020) was used. The user profile locations and parsed text addresses were then geocoded using Google geocoding (API). Using these data, three analyses were conducted: geolocating agreement, comparison of Twitter Ratio and Sentiment indexes, and specific disaster-related content analysis.

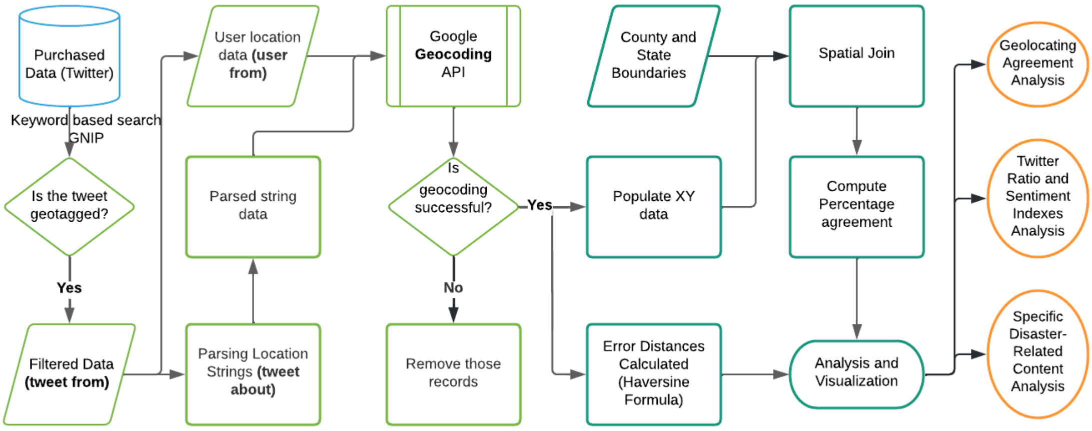

Purchased Data (Twitter)
Keyword based search GNIP
Is the tweet geotagged?
Yes
Filtered Data (tweet from)
User location data (user from)
Parsed string data
Parsing Location Strings (tweet about)
Google Geocoding API
Is geocoding successful?
Yes
No
Remove those records
County and State Boundaries
Populate XY data
Error Distances Calculated (Haversine Formula)
Spatial Join
Compute Percentage agreement
Analysis and Visualization
Geolocating Agreement Analysis
Twitter Ratio and Sentiment Indexes Analysis
Specific Disaster-Related Content Analysis

Fig 1: Research Methodology

For **Geolocating Agreement Analysis**, there were two approaches undertaken; one with the calculation of displacement distance from geotagged location and the other, checking whether the original administrative unit for that tweet got changed. The error distance was calculated using the Haversine formula. For percentage agreement, these tweets were checked if both locators lied within the same administrative boundaries thereby calculating the accuracy based on geotagged location. These results were then visualized.

For **Twitter Ratio and Sentiment Indexes Analysis**, in the discussion intensity, the formula used was tweets discussing disaster divided by total number of tweets from each county. The county twitter ratio scores using the three different methods of geolocating data were compared for accuracy. The sentiment analysis score was evaluated through VADER (Hutto et al. 2014) sentiment analysis tool, which produces sentiment scores from -1 to +1. These sentiment scores too were aggregated by county.

For **Specific Disaster-Related Content Analysis**, the tweets were filtered checking through commonly used keywords for such purposes. These tweets are then checked for displacement by county for the affected state, viz. we have considered the case in the state of Texas here. These were then visualized for analysis.

Results

Figure 2 shows the agreements of predicting users' locations through 'user from' and 'tweet about' addresses at the state level. The state of Texas has the highest sample space since it is one of the prime affected zones and shows highest accuracy. Whilst Louisiana, although being at the front-end of the affected states, lags slightly behind. This goes on to show that closer to impact zones, locators would be interchangeable – however if one moves further away, the interchangeability will depend on the threshold allowable for the study while using this map as a reference (Figure 2). For veracity, only state analysis results are shown. The maximum frequency is present in the 50-60 percentage agreement and has a relatively normal distribution of data. Similar level analyses were done as shown in the methodology for the other three and reported in the later section.

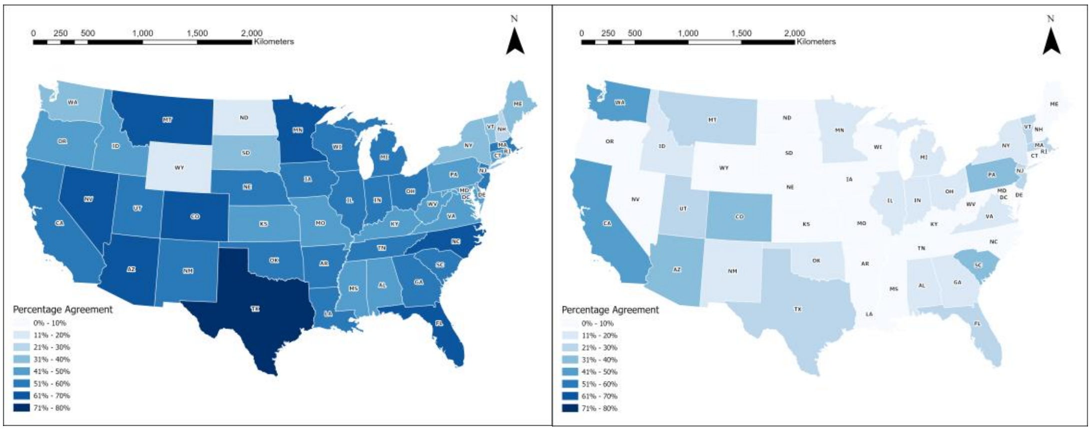


Fig 2: Percentage Agreement in State level for "user from" and "tweet about"

For state and county levels, accuracy drops sharply from states to counties. While the states fare relatively better, its performance does seem little worse than the high accuracy in country levels. Country based accuracy is the highest in the chart (Figure 3), followed by State, County, Block, 1km, 30m and a perfect match. The last three levels are based on the error distances. This chart gives the entire view how the whole dataset's accuracy fluctuates for given allowable errors. It shows how sharply the accuracies increase with increasing administrative boundary sizes as compared to the flatline on the lower end. It goes on to show that at lower levels, the displacement will have a lower limit but because of the low allowance of error, it will not increase either. Comparatively however the location details from the 'tweet about' locations did much worse in terms of accuracy. The chart (Figure 3) shows the overall accuracy trend that occurs as we enlarge the lowest administrative unit considered from county to state to country. As even at state level the accuracy percentage is so incredibly low, it cannot be recommended to use this method for any county level operations.

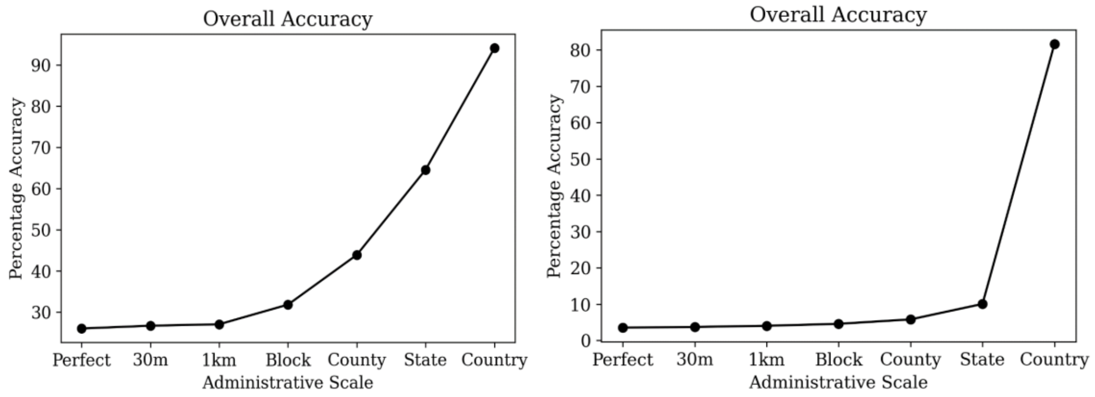


Fig 3: Overall Accuracy for "user from" and "tweet about"

Figure 4 showcases percentage agreements of geocoding in separate phases of the disaster. It reiterates the concept that at perfect match to 30m and 1 km error distances the percentage agreements are similar - while it increases for the better as the area accounted for increases in administrative standards. If the ridges are noticed, it becomes clear that at higher levels it is much more plausible to use them interchangeably. It would be highly risky to do so in the county level during response period as it may cause serious misinformation and thereby sabotage the rescue process rather than enhancing it.

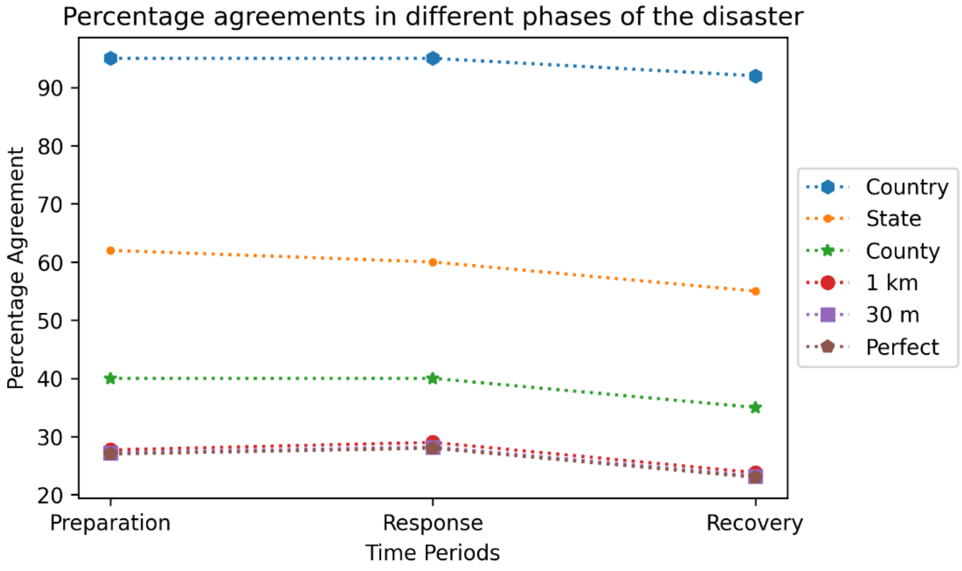


Fig 4: Percentage Agreements in separate phases of disaster for "user from" data

Discussion and Conclusion

The results for twitter ratio have high correlation from the control to testing set. This is possibly because the areas with higher frequency of tweeting the event would have users active enough to update their profile or put proper context in the tweets discussing the disaster event. The resultant maps are shown below (Figure 5). This high similarity indicates at this usage the locations obtained maybe substituted even at county level.

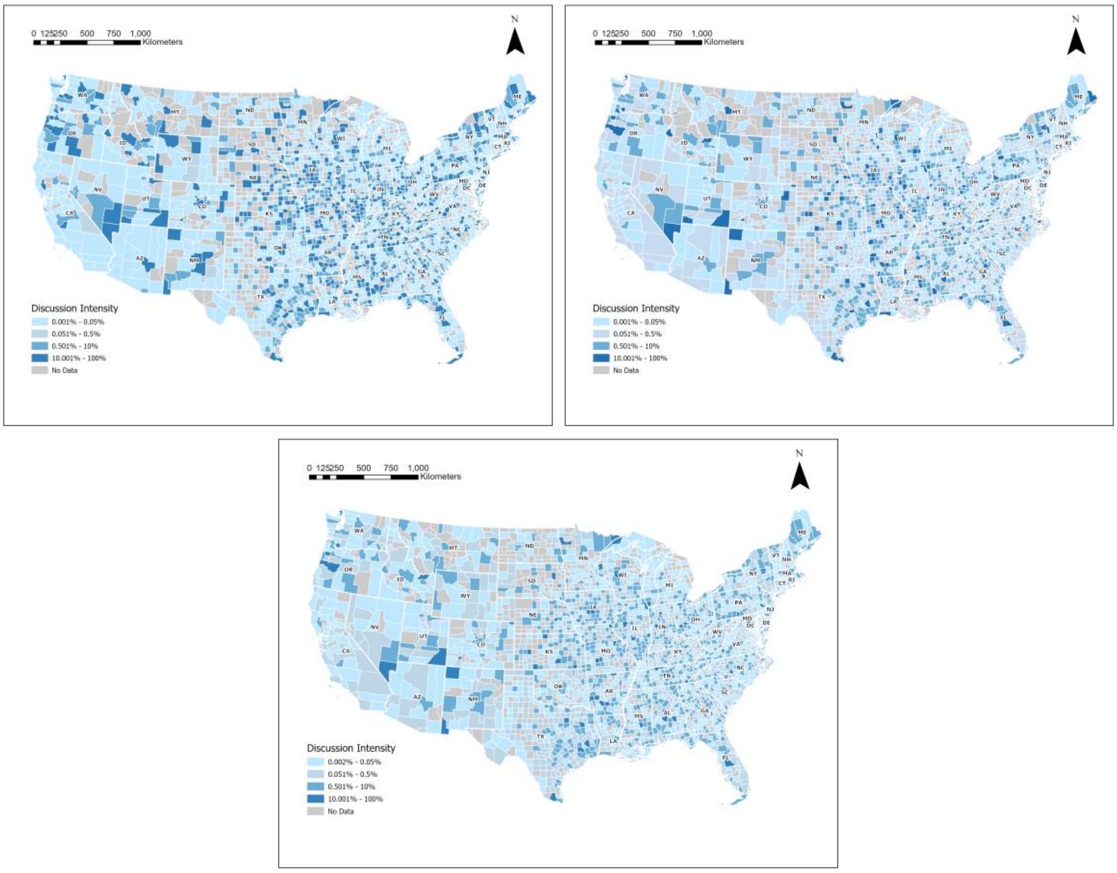
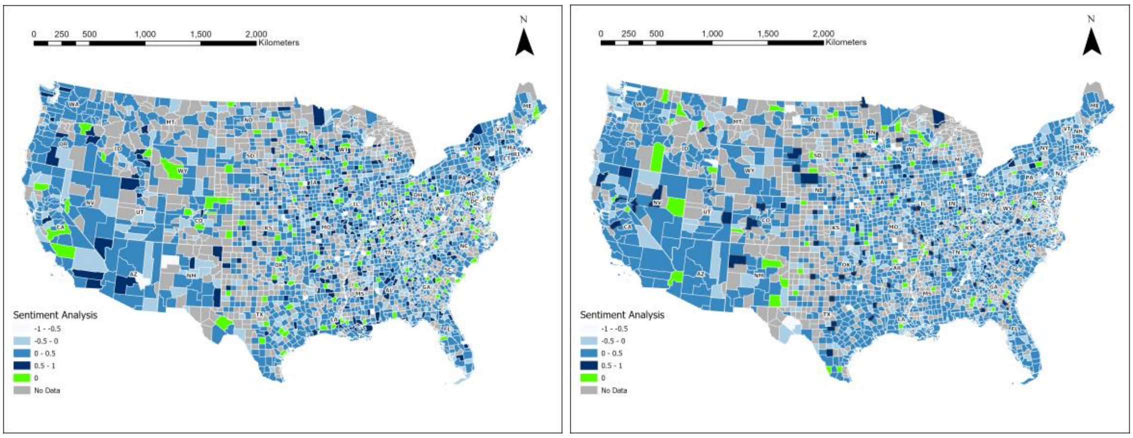


Fig 5: Twitter Ratio for "tweet about", "tweet from" and "user from"

Sentiment Analysis on the other hand had very low correlation from the geocoded location to the geotagged one. This has been so since the changes in highly emotional tweet's location would be enough to have a big impact on the aggregate sentiment score of a county (Figure 6). The disparities in the pattern belie any hopes of using these locations interchangeably.



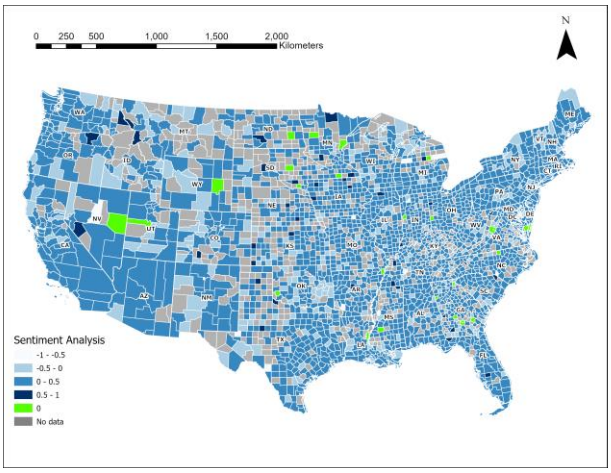


Fig 6: Sentiment Analysis for "tweet about", "tweet from" and "user from"

Findings from this project show that data from social media used during disasters can be used depending on the locational constraints and threshold limit of errors. It would be subject to each unique study type– while this study shows a keystone that can be referenced to and based upon while making such decisions.

References

Cinelli, M., Cresci, S., Galeazzi, A., Quattrociocchi, W., & Tesconi, M. (2020). The limited reach of fake news on Twitter during 2019 European elections. PLOS ONE, 15(6), e0234689. https://doi.org/10.1371/journal.pone.0234689

Dwoskin, E. (2014, April 8). The Race To Locate Twitter Users. WSJ. https://www.wsj.com/articles/BL-DGB-34180

Graham, Mark, Scott A. Hale, and Devin Gaffney. 2014. "Where in the world are you? geolocation and language identification in Twitter." *The Professional Geographer* 66 (4): 568-578. 10.1080/00330124.2014.907699.

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. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Kaushik Soni, locationtagger, (2020), GitHub repository, https://github.com/kaushiksoni10/locationtagger

Lin, Binbin, Lei Zou, Nick Duffield, Ali Mostafavi, Heng Cai, Bing Zhou, Jian Tao, Mingzheng Yang, Debayan Mandal, and Joynal Abedin. n.d. "Revealing the Global Linguistic and Geographical Disparities of Public Awareness to Covid-19 Outbreak through Social Media." NASA/ADS. Accessed March 4, 2022. https://ui.adsabs.harvard.edu/abs/2021arXiv211103446L/abstract.

Liu, Xiaomo, Quanzhi Li, Armineh Nourbakhsh, Rui Fang, Merine Thomas, Kajsa Anderson, Russ Kociuba, et al. 2016. "Reuters Tracer." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 10.1145/2983323.2983363.

Tsao, S. F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., & Butt, Z. A. (2021). What social media told us in the time of COVID-19: a scoping review. The Lancet Digital Health, 3(3), e175–e194. https://doi.org/10.1016/s2589-7500(20)30315-0

Wang, Y., Wang, T., Ye, X., Zhu, J., & Lee, J. (2015). Using social media for Emergency Response and Urban Sustainability: A Case Study of the 2012 Beijing Rainstorm. Sustainability, 8(1), 25. https://doi.org/10.3390/su8010025

Wong, J., & Wong, N. (2021). The economics and accounting for COVID-19 wage subsidy and other government grants. Pacific Accounting Review, 33(2), 199–211. https://doi.org/10.1108/par-10-2020-0189

Zhou, B., Zou, L., Mostafavi, A., Lin, B., Yang, M., Gharaibeh, N., Cai, H., Abedin, J., & Mandal, D. (2022). VictimFinder: Harvesting rescue requests in disaster response from social media with BERT. Computers, Environment and Urban Systems, 95, 101824. https://doi.org/10.1016/j.compenvurbsys.2022.101824

Zou, Lei, Nina S. Lam, Heng Cai, and Yi Qiang. 2018. “Mining Twitter Data for Improved Understanding of Disaster Resilience.” Annals of the American Association of Geographers 108 (5): 1422-1441. 10.1080/24694452.2017.1421897.

Zou, Lei, Nina S. Lam, Shayan Shams, Heng Cai, Michelle A. Meyer, Seungwon Yang, Kisung Lee, Seung-Jong Park & Margaret A. Reams (2019) Social and geographical disparities in Twitter use during Hurricane Harvey, International Journal of Digital Earth, 12:11, 1300-1318, DOI: 10.1080/17538947.2018.1545878