SOCIAL MEDIA ATTENTION CONCENTRATES IN POPULOUS AREAS DURING DISASTERS

An Undergraduate Research Scholars Thesis

by

MIGUEL ESPARZA

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Ali Mostafavi

May 2020

Major: Civil Engineering

TABLE OF CONTENTS

ABSTRACT

Social Media Attention Concentrates in Populous Areas
During Disasters

Miguel Esparza
Department of Civil Engineering
Texas A&M University

Research Advisor: Dr. Ali Mostafavi
Department of Civil Engineering
Texas A&M University

The objective of this study is to examine and quantify the relationships among sociodemographic factors, damage claims and social media attention on areas in disasters. People seek situational awareness during disasters in order to perceive the risks and cope with community disruptions. With the increased use of social platforms, social media has become an important communication channel for people to share and seek situational information and support disaster response. Recent studies in disaster informatics have recognized the presence of bias in the representation of social media activity in different areas affected by disasters. To explore related factors for such bias, geo-tagged tweets have been used to study the extent of social media activity in disaster-affected areas to evaluate whether vulnerable populations remain silent on social media. However, less than 1% of all tweets are actually geo-tagged, therefore attempts to understand the representative of geotagged tweets to the general population have shown that certain populations are over or underrepresented. To overcome this limitation, incorporating relevant tweets identified based on their content is essential. Here, we conducted a content-based analysis to filter the tweets related to super-neighborhoods in Houston during

Hurricane Harvey and cities in North Carolina during Hurricane Florence. By examining the relationships among sociodemographic factors, the number of damage claims and the volume of tweets, we find that social media attention concentrates in populous areas and is independent of education, language, unemployment, and median income. The relationship between population and social media attention is characterized by a sub-linear power law, indicating a large variation among the sparsely-populated areas. Using a machine learning model to label the topics of the tweets, we show that social media users pay more attention to rescue and donation related information, but the topic variation is consistent across areas with different levels of attention. These findings contribute to a better understanding of the spatial concentration of social media attention regarding posting and spreading situational information in disasters, and suggest planners and policymakers to better use social media data for equal disaster treatment.

ACKNOWLEDGMENTS

This research would not have been possible without the help of the following mentors at the Urban Resilience, Network, and Informatics Lab: Chao Fan, Jennifer Dargin, Fangsheng Wu, Bora Oztekin, and Dr.Ali Mostafavi. Thank you all for your guidance and support throughout the course of this research project

This material is based in part upon work supported by the National Science Foundation under Grant Number CMMI-1846069 (CAREER) and the Amazon Web Services (AWS) Machine Learning Award. The authors also would like to acknowledge the funding support from the National Academies' Gulf Research Program Early-Career Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation and Amazon Web Services.

CHAPTER I

INTRODUCTION

Equitable response and recovery are important for reducing causalities and property losses in areas unevenly impacted by a disaster. One key approach to equitable disaster response is to be aware of the situations and disparities of the impacts of disasters among different affected areas. Hence, enhancing equitable situational awareness that can aid in adapting to uneven disaster impacts is critical (Seppänen & Virrantaus, 2015).

With the increased use of digital devices and social platforms, social media offers the possibility of improved disaster communication. Unlike traditional media such as radio and news articles, social media enables augmented information capacity and rapid interactivity, which allow people to post and share situational information more efficiently (Zhang, Fan, Yao, Hu, & Mostafavi, 2019). This makes social media data attractive for large-scale laboratories conducting disaster research (Ghani, Hamid, Targio Hashem, & Ahmed, 2018), including detecting disruptive events (Chao Fan & Mostafavi, 2019; Chao Fan, Mostafavi, Gupta, & Zhang, 2018), characterizing information propagation (Sutton et al., 2015; Yang et al., 2019), and measuring human sentiments (Ragini, Anand, & Bhaskar, 2018) during disasters. In addition, Twitter has a feature of geotagging that allows tweets to be associated with accurate longitudes and latitudes. This enables researchers to know where the users were and where the tweets were generated. Due to this benefit, a vast number of studies have analyzed the geotagged tweets for various objectives, such as estimating disaster scales (Kumar, Hu, & Liu, 2014), mapping disaster situations (Z. Li, Wang, Emrich, & Guo, 2018), and quantifying human movement (Wang, 2015). These studies make social media well-supported for situational awareness and disaster management.

Despite the opportunities borne by social media, it may lead to issues of representativeness of the disaster situation. Existing studies show that social media users pay various levels of attention on different areas in disasters (Malik, Lamba, Nakos, & Pfeffer, 2015). That is because the percentages of people using social media and households without telephone services vary significantly by location (Xiao, Huang, & Wu, 2015). Recognized and emphasized by existing research, studies on the variation of social media attention in terms of the social and geographical disparities of different areas are emerging (Zou et al., 2018). For example, Madianou revealed that low-income participants have diminished social media opportunities, which lead to a deepening of social inequalities on social media and disaster recovery (Madianou, 2015). Xiao et al. examined the spatial heterogeneity in the generation of geotagged tweets and found that socioeconomic factors are important in predicting the number of tweets in census tracts (Xiao et al., 2015). Kryvasheyeu et al. found relationships between the proximity to a hurricane's path and the quantity of hurricane-related geotagged tweets (Kryvasheyeu et al., 2016). These studies indicate the importance of sociodemographic and damage factors in influencing social media attention on different areas through the use of geotagged tweets.

As documented in existing studies, however, fewer than 0.42% of all tweets are associated with accurate geospatial information such as latitudes and longitudes (Cheng, Caverlee, & Lee, 2010), and only about 1 to 1.5% of the tweets are geo-coded with cities or neighborhoods (Morstatter, Pfeffer, Liu, & Carley, 2013). The low representation is because the majority of the users disable the geotagging function to protect their privacy (Malik et al., 2015).

Hence, a mass of disaster-related tweets is proactively ignored in existing studies. In addition, the situation provided in the texts of the tweets may not match the geotags where users generated the tweets (C Fan, Wu, & Mostafavi, 2020). Moreover, trending topics of the information that people care about on social media also have spatial patterns, which correspond to the disaster footprint (Resch, Usländer, & Havas, 2018). However, existing studies only focus on Twitter activities, and the topic of the information attracting various levels of users' attention is usually overlooked. All these problems in existing research would lead to biases in claiming the relationship between geographical heterogeneity and social media attention, which may exacerbate the inequality of response actions and resource allocation in disasters.

To overcome the issues in geotagged tweets and activity-only analysis, there is a real need for content-based analysis to examine the disparities of social media attention on different areas. To this end, this study aims to examine the influence of sociodemographic factors and damages on social media attention during disasters by identifying the location-relevant tweets from their content. Three research questions will be the guide of this study:

- What sociodemographic factors affect social media attention on areas in disasters?
- How can we quantify the relationship between social media attention and population or damage claims? Is the relationship consistent at different levels of scale?
- What information topics do users pay more attention to on social media? Does is differ from areas with different levels of social media attention?

To achieve such an objective, we filtered relevant tweets based on the appearance of disaster-affected super-neighborhoods and cities in the contents of the tweets, quantified the relationship between social media attention and the sociodemographic factors of areas using regression models, and examined the variation of the topics with different levels of user attention. The super-neighborhoods in Houston during Hurricane Harvey and cities in North Carolina during Hurricane Florence are both included in this study to uncover the cross-hazard similarity and cross-scale adaptability of social media attention patterns. Assessing these patterns regarding geospatial disparities through the content-based analysis will provide an understanding of the inequalities that may cause unfair disaster treatment.

CHAPTER II

RELATED WORK

A number of prior studies have explored the geospatial patterns of social media activities, investigated social media use among different groups of people, and developed techniques for classifying tweets with humanitarian topics (Fischer-Preßler, Schwemmer, & Fischbach, 2019). This section will discuss these prior research works and identify the opportunities and necessities of conducting this study.

The geotags of the tweets enable researchers to capture the specific locations where the tweets were created (Sloan & Morgan, 2015). Understanding the geospatial patterns of the volume of tweets, to some degree, can imply the population activities or disaster situations. Hence, the geospatial patterns of social media data have been extensively investigated using geotagged tweets in recent disasters such as hurricanes (MacEachren et al., 2010), flooding, fires and haze (Kibanov, Stumme, Amin, & Lee, 2017). One stream of the studies focuses on mapping and assessing the damages for disaster-affected areas using geotagged tweets. For example, Middleton et al. proposed a social media disaster mapping platform to geo-parse the real-time data streams and assess the disaster impacts (Middleton, Middleton, & Modafferi, 2014). To estimate the scale of the disasters, predictive and approximate methods including kernel density estimation (Chao Fan, Jiang, & Mostafavi, 2020) and network prediction model (Rahimi, Cohn, & Baldwin, 2015) are adopted in analyzing the geographical distribution of geotagged tweets. As researchers realize the limitations of geotagged tweets (Malik et al., 2015), recent studies are trying to uncover the underlying factors that induce the biases in geotagged tweets. The primary factors examined in existing research are demographic and socioeconomic factors (Hecht &

Stephens, 2014). Prior studies show that there is a positive association between populations with high-level academic degrees and the number of geotagged tweets (Jiang, Li, & Ye, 2019). Although existing studies have shown the influence of social media activities, the geotagged tweets only account for about 1% of the total volume of tweets, and are not representative of the entire social media activities in disasters. In addition, it is also a case that the geotags of the tweets do not match the content of the tweets. People might post a geotagged tweet in a location different from the location where an event was described in the texts. Hence, it is necessary to analyze the social media attention by parsing the content of the tweets.

Sociodemographic factors such as education, minority, and median income usually affect not only the preparedness and response to the disasters, but also the accessibility and use patterns of social media in the affected areas. Specifically, as shown in the results of large-scale surveys, education can reveal people's ability to understand information about emergency plans or warning information to avoid dangerous situations (Cutter, Boruff, & Shirley, 2003). The language gap in a disaster creates cultural barriers in a community that can exacerbate the disaster response for people who do not speak the native language well (Frigerio & De Amicis, 2016). The minority status (i.e., nonwhites) or a lack of wealth in the population contributes to social vulnerability through the lack of access to resources and internet during and after disasters (Cutter et al., 2003). In addition, existing studies have revealed that racial/ethnic and health status–related disparities exist in Internet access, but not significantly affect social media use patterns (Chou, Hunt, Beckjord, Moser, & Hesse, 2009). People with different demographic and psychosocial background also have different perceptions of using social media (Keating, Hendy, & Can, 2016). In particular, Neubaum et al. found that the motivations of people using social media in disasters is to share the emotions and purse empathic concerns (Neubaum, Rösner, Rosenthal-Von Der Pütten, & Krämer, 2014). Given this fact, the population size tends to strongly influence the social media attention on an area. Recent studies help us identify multiple potential factors that might influence the social media attention, which provide empirical and theoretical evidence for selecting variables in this study. Despite the progress made in capturing the relationships between social media use and sociodemographic factors, the relation between the social media attention in an actual disaster event and its related factors still remains unknown.

Finally, the topics of the information delivered by the tweets have also attracted significant public and research interest. Specifying the topics of the tweets would help residents and first responders better understand the bias of social media attention regarding the topics of the information, which is absent in existing studies. This will be achieved by adopting the techniques for batch labeling of social media data. To do this, computer science researchers have developed multiple advanced techniques for classifying tweets with humanitarian categories (Imran, Castillo, Diaz, & Vieweg, 2014). The machine learning approaches for disaster applications started with unsupervised learning such as Latent Dirichlet allocation topic modeling for detecting the trending topics in disasters (Chao Fan et al., 2018; Hidayatullah, Aditya, Karimah, & Gardini, 2019). While these approaches can capture the topics of massive tweets in near real-time, the outputs are not stable due to excessive noise present in tweets. With the development of supervised learning, multiple advanced learning approaches such as Naïve Bayes classifier (Hutto & Gilbert, 2014) were adopted in identifying the topics of the tweets. These approaches usually trained the model on labeled data and implemented on the datasets from similar hazards or crises. For example, Li et al. proposed a domain adaptation approach integrating with the Naïve Bayes classifier to label the social media data in emerging target disasters (H. Li, Caragea, Caragea, & Herndon, 2018). Caragea et al. presented an approach based on Convolutional Neural Networks to identify informative messages in social media streams in disasters (Caragea, Silvescu, & Tapia, 2016). A recent study which fine-tuned a BERT-based classifier further enhanced the performance of the deep learning models for categorizing the tweets (C Fan et al., 2020). The advancement of machine learning approaches provides us with unique opportunities of analyzing the distribution of social media attention regarding the topics of the information.

In summary, recent work not only offers the techniques and empirical evidence for analyzing social media data in disasters, but also points out the necessity of examining the inequality in social media contents regarding the volume and topics of the posts. On the basis of these related work, the rest of the paper will develop a conceptual model and apply statistical tests to quantify the relationships among social media attention, sociodemographic factors, damage claims, and information topics.

CHAPTER III

CONCEPTUAL MODEL

We base the literature review on the opportunity and necessity of examining the relationships between social media attention and socio-demographic factors by looking into the content of the tweets. There are plenty of sociodemographic factors that can characterize the societal attributes of the areas in disasters. As suggested by existing studies discussed in chapter 2, sociodemographic factors such as population, percentage of people with high-school degrees, percentage of people whose native language is English, percentage of unemployment, percentage of minority and median income tend to have an influence on attracting the attention of social media users (Jiang et al., 2019). Hence, this study mainly focuses on these factors and their relationships with social media attention. Despite potential influence of these factors, according to the psychological evidence (Neubaum et al., 2014), social media attention and activities are motivated by acquiring empathic concerns from people in the same situation. Due to the localized impacts of disasters, the people from the same area tend to be in the same situation. Assuming the number of social media users is proportional to the population size, social media attention might be driven by the population of an affected area. As such, we hypothesize that:

H1. Population size and volume of damage claims have a strong, positive relationship with social media attention in areas during disasters, and social media attention is independent of other sociodemographic factors.

Despite recent advances in uncovering a significant correlation between social media attention and sociodemographic factors, many applications from pandemic prediction to disaster response, require a quantitative understanding of how the relationship can be used to predict the damages or sociodemographic attributes. In addition, the variation in both population and the number of tweets is extremely large among different areas. The variation might be further amplified by expanding the scale of the areas (from super-neighborhood scale to city scale). Thus, linear models in linear space might not be able to capture the relationship very well. Logarithmic space is commonly used to shrink the scales in the data, which respond to the skewness toward large variations in the dataset. Meanwhile, the percentage changes or multiplicative factors can also be identified from the data in logarithmic space. Then, linear models can be applied to the logarithmic space to quantify the relationship. This will lead to an examination of a power-law relationships between two variables. Therefore, we hypothesize that:

H2a. Social media attention and population (and damage claims) follows a power-law relationship of the form $y \approx x^\beta$, where x is the population of a city or super-neighborhood, y is the number of tweets, and β is the scaling exponent.

H2b. The power-law relationship is consistent at both city and super-neighborhood scales.

Social media platforms have no limitations on the topics of the disaster information posted by users. In previous disasters, the situational information included infrastructure and utility damages, affected individuals, rescue and volunteering efforts. This information could raise attention of the first responders, relief organizations, and residents outside the affected areas. By analyzing the content of the tweets, a prior study shows that social media becomes an important tool for communicating relief efforts in different areas (C Fan et al., 2020). For example, relief organizations shared the information about their resources and locations, and people at risk posted their needs to connect to the relief organizations. According to this empirical evidence, we hypothesize that:

H3a. Social media users pay more attention to rescue, volunteering, and donation-related information.

H3b. The topics that users pay more attention to do not differ from areas with different levels of social media attention.

Based on the aforementioned constructs, we propose a conceptual model for organizing the relationships among multiple variables and developing hypotheses (Figure 1).

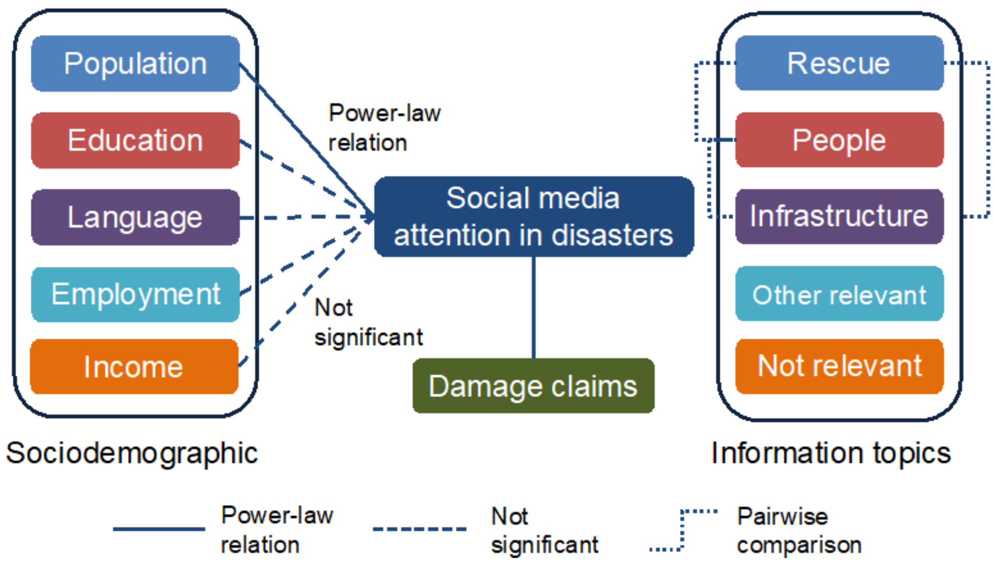


Figure 1. Conceptual model for examining social media attention in disasters.

CHAPTER IV

METHODOLOGY

To test the hypotheses and answer the research questions, we conducted case studies on super-neighborhoods in Houston during Hurricane Harvey and cities in North Carolina during Hurricane Florence. Hurricane Harvey, a category 4 storm, made landfall on August 27, 2017, in Houston. The torrential rainfall that occurred during Harvey and the slow movement of the storm system led to intense flooding in Harris County, which caused the destruction of 50,000 homes (Pulcinella, Winguth, Allen, & Dasa Gangadhar, 2019). Since the majority of the damages occurred in Houston during Harvey, we determined to use super-neighborhoods in Houston in this case. Hurricane Florence formed in August 2018, and the early warning was sent about September 6. On September 14, Florence struck the southeastern coast of North Carolina and caused 53 fatalities in three states (NC, SC, and VA) and 16-40 billion in damage, where 50% of the damage projected as uninsured losses due to residential flooding (Paul, Ghebreyesus, & Sharif, 2019). Hurricane Florence made a larger scale of impacts on multiple cities in North Carolina. Hence, we selected cities affected by Florence in this case study.

Seven attributes were used to indicate attention disparities on social media during the disaster events. In this study, census data from American Factfinder (American FactFinder, 2020) was collected for the following social groups: population size, the percentage of the educated population, the percentage of people who are unemployed, the percentage of people whose native language is not English, the percentage of minority groups, and the medium income. These social groups were observed in 84 super neighborhoods for Hurricane Harvey in Houston, and 57 cities in North Carolina affected by Hurricane Florence.

To assess the physical damage among populations, this study uses the number of Federal Emergency Management Agency (FEMA) claims in each area of study. The damage claims were filed by people in a disaster when their properties or services were disrupted. This data was collected using HydroShare, which enables storage, management, sharing, publication, and annotation of data associated with hydrological studies (Horsburgh et al., 2016). FEMA claims are an effective way to measure the physical damage of a disaster because claims are used to compensate the public's property damage.

After these attributes were filtered by removing outliers with insignificant data, the descriptive statistics were computed and are summarized in Table 1. Each attribute provides unique information about the social composition of the areas of study, as highlighted by Table 1. For example, Language (%) and Minority (%) have similar standard deviations, but their means do not have a similar relationship. The large disparities between these two averages imply that cultural barriers between minority groups and language groups have a different degree of impact between cities.

Table 1. Summary Statistics for Sociodemographic Data and FEMA Damage Claims.

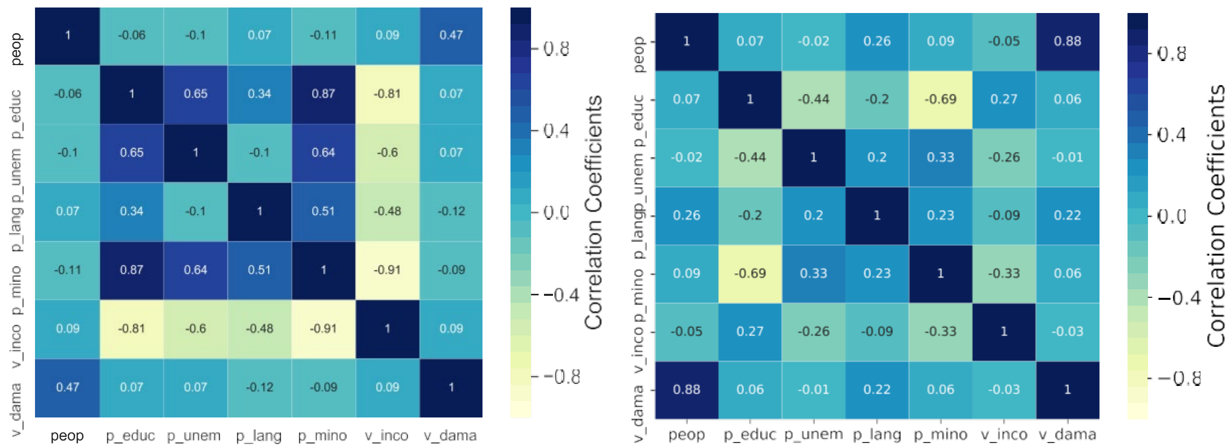
Variables	Hurricane Harvey					Hurricane Florence				
	Obs.	Mean	SD	Min	Max	Obs.	Mean	SD	Min	Max
Population size	84	25,955	20,863	2,031	119,598	57	6,824	17,007	112	106,476
Education (%)	84	23.3	9.5	2	40	57	89.3	7.3	67.7	100
Unemployment (%)	84	7.9	4.0	2	23	57	5.4	3.6	0	19.6
Language (%)	84	47.2	21.7	10	89	57	5.6	4.9	0	25.9
Minority (%)	84	76.3	22.6	23	99	57	21.9	17.9	1.0	61.0
Median Income ($)	84	57,029	27,137	25,489	152,092	57	64,940	37,338	2,453	160,311
Damage claims	84	626	3,102	7	28,579	57	1,141	1,916	6	10,821

In this study, we define social media attention mainly based on the total number of tweets which integrates all types of activities of the users. The relationships between sociodemographic factors (and damage claims) and each type of tweets will be quantified as well. To examine the social media attention on different areas, we collected 2 million tweets for both disaster events. The time period for collecting tweets for Hurricane Harvey spans from August 26 to September 4, which covers the response and recovery phases of Hurricane Harvey. Since Hurricane Florence was a powerful and long-lived disaster event, data was collected from September 6 until September 26. The dataset includes all of the tweets that were posted by the users whose profiles show the locality of Houston or North Carolina, and the tweets that were geotagged in our predefined bounding boxes which covered all disaster-affected areas. Then, we manually identified the names, abbreviations, main buildings, key roads, and relevant keywords for super-neighborhoods in Houston so that we could filter the relevant tweets for each super-neighborhood. Cities in Florence are much larger in scale than super-neighborhoods and include a vast number of buildings and roads. Hence, only the names of the cities were employed for filtering relevant tweets. The descriptive statistics of the Twitter data for super-neighborhoods and cities are summarized in Table 2. As shown in the table, apparently, retweets account for the largest proportion of the social media attention with largest variations and maximum values in both super-neighborhood and city scales, and replies and quotes account for a very small proportion in the total amount of tweets.

Table 2. Summary Statistics for Twitter data.

Variables	Hurricane Harvey					Hurricane Florence				
	Obs.	Mean	SD	Min	Max	Obs.	Mean	SD	Min	Max
Original tweets	84	99.5	155.4	0	671	57	376.5	1,194.8	0	8,354
Retweets	84	374.7	674.5	0	3,165	57	2,968.8	7,860.8	0	37,659
Replies	84	14.3	23.0	0	136	57	82.5	254.2	0	1,508
Quotes	84	18.5	36.9	0	214	57	50.0	150.3	0	968
Total tweets	84	477	815	0	3,721	57	3,395	9,148	0	48,198

Pearson's correlation coefficients were calculated for each pair of variables (Figure 2) to confirm the independence among the selected variables in the analysis. The results show that the number of damage claims is strongly and positively proportional to the population size, while other factors such as median income, unemployment, minority, language, education are all independent of the population and damage claims. This finding can be observed in both super-neighborhood and city scales. However, the relationships among education, unemployment, language, minority and medium income are rather different in different scales. The results indicate that, to some degree, these variables are correlated with each other. This signifies that we cannot include these variables in a single regression model due to their dependencies. This observation indicate that we have to conduct pair-wise comparison between social media attention and social variables.



(a) (b)

Figure 2. Correlation analysis for sociodemographic.

The advances in machine learning techniques enable automatic labeling of topics for tweets, which allows us to examine what type of information social media users pay more attention to. Existing studies have developed high-performance machine learning models, in particular, for classifying disaster-related tweets. We adopted a recently published work (C Fan et al., 2020) in which an advanced BERT-based model is trained and tested for labeling hurricane-specific tweets with humanitarian categories. The BERT-based model is designed to train bidirectional representations that embed the words, position of the words, and the segments of a tweets. To implement the model, we first employed the definition of the humanitarian categories (Alam, Ofli, & Imran, 2018). Table 3 shows the example tweets in Hurricane Harvey and Florence for each humanitarian topics. There are five topics covering the information from infrastructure damages, affected individuals, to rescue efforts. In addition to these three specific topics, the other category (i.e., other relevant information) can also be identified, in which the tweets are related to the disaster events but may not deliver specific information related to the three main categories. Hence, in analyzing the social media attention, the tweets in the "other relevant" category is still considered.

The analysis on the discrepancies of social media attention regarding information topics will mainly focus on the three main categories: infrastructure damages, affected individuals, and rescue efforts (Table 3). After defining these humanitarian topics, the adopted BERT-based machine learning model is applied to the filtered tweets for all selected super-neighborhoods in Houston and impacted cities in North Carolina. The model will generate measures of the closeness of the tweets to each humanitarian category, represented by a probability. The output of the model is the label for each tweet based on the highest probability of the category. By doing so, we will compile the number of tweets in each topic for each area (super-neighborhoods and cities).

Table 3. Example Tweets for Different Humanitarian Topics in Hurricane Harvey and Florence.

Topics	Example tweets
Infrastructure and utility damages	*"Streets in downtown are filling with a lot of water – pls don't try to drive right now #Houston #Harvey* *"I-40 is flooded at Burgaw and at Castle Hayne. "We don't have a land access to Wilmington."*
Affected and injured individuals	*"I love over three hours from the North Carolina beaches but saw people from the coast in stores today buying emergency supplies"* *"my sister is at the house. No word on her apt downtown but we think it probably flooded"*
Rescue, volunteering, or donation effort	"*Night shift volunteers also sought at downtown Brown convention center shelter. #Harvey2017* *"Beulah Baptist Church in Calabash is taking donations for storm victims. They are in serious need of dog and cat food."*
Other relevant information	"find open restaurants and details on flooded areas at Downtown Houston" "*Hurricane #Florence rain totals so far. Oriental 21.6 Surf City: 16.6 New Bern: 14.26 Swansboro: 14.25 Calabash: 12…"*

CHAPTER V

RESULTS

As mentioned previously, the variation of social media attention is extremely large among different geographic areas (super-neighborhoods or cities). To make the comparison straightforward, we intuitively divided the super-neighborhoods and cities respectively into two groups with different levels of social media attention. In this study, we set the threshold for social media attention to be 300 tweets for both hurricanes to balance the size of two groups of areas in both cases. Then, we conducted a two-sample t test for difference of means in two groups for each hurricane.

Figure 3 shows the differences of population (Figure 3a and 3c) and damage claims (Figure 3b and 3d) in groups of areas with high or low social media attention. In the case of Hurricane Harvey, the population mean in areas with high social media attention is about 35,000, while the population mean in areas with low attention is only 2,000. For damages, since it is correlated to the population size, the mean of damage claims in areas with high social media attention is about 3,000. The areas with low social media attention only have an average of 800 damage claims. In addition, population size within a standard deviation in the group of areas with high attention is always greater than that of areas with low attention. We can observe these patterns in city scales during Hurricane Florence. Hence, we can claim that areas with high social media attention have much larger populations and a greater number of damage claims than those of areas with low social media attention. Through the test of significance, this finding is significant and consistent in both super-neighborhood and city scales. Since the number of damage claims is correlated to the population of the areas, we can claim that the social media attention is concentrated in populous areas.

The focus of social media attention might also be affected by the composition of the population. However, as shown in Table 4 below, the means of the two groups in the areas are close to each other for all variables. This result implies that the composition of the population in terms of any other sociodemographic factors (i.e., education, unemployment, language, minority and median income) in the two groups is not significantly different. The pattern can be identified from both super-neighborhoods and cities. Hence, social media attention is independent of these sociodemographic factors, and only influenced by population size and damage claims. Therefore, we can accept hypothesis **H1**.

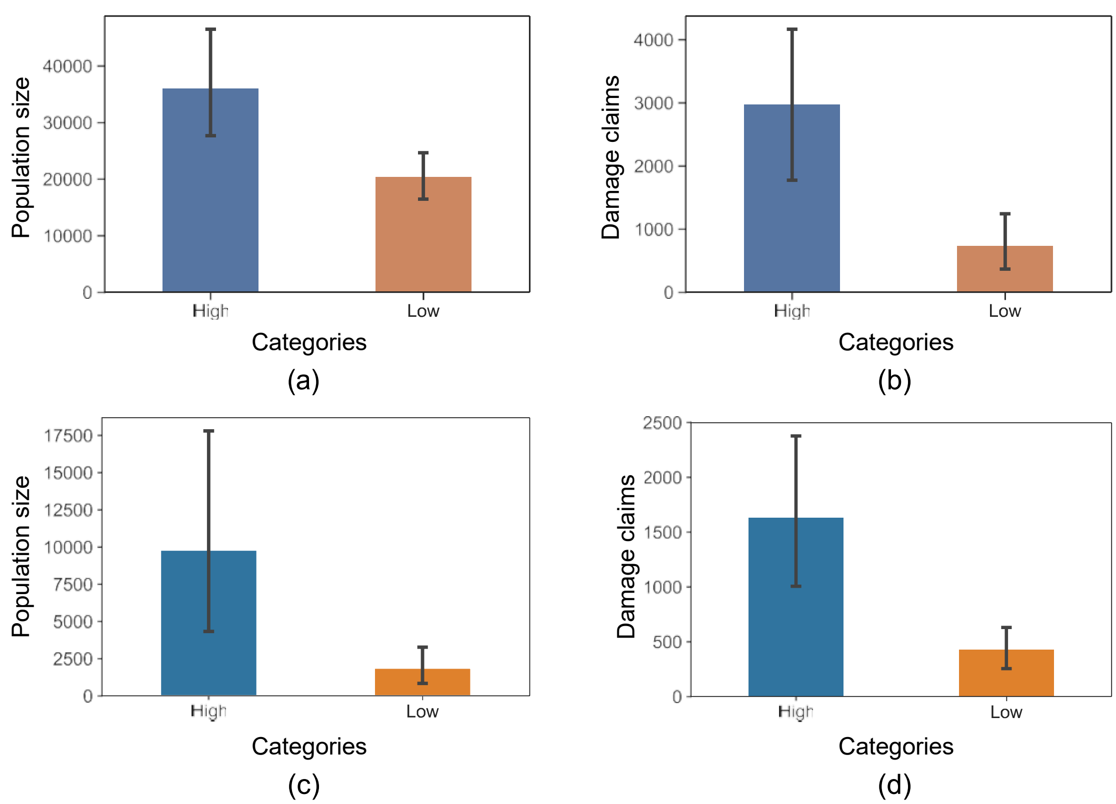


Figure 3. Population and number of damage claims in areas with different level of attention.

Table 4. The differences and significance of sociodemographic variables in areas with different level of attentions.

Cases	Variables	Mean of high attention areas	Mean of low attention areas	*P-value*
Super-neighborhoods in Hurricane Harvey	Education (%)	21.8	24.8	Not significant
	Unemployment (%)	7.5	8.1	Not significant
	Language (%)	44.4	48.7	Not significant
	Minority (%)	71.2	80.0	Not significant
	Medium income ($)	63,459	52,255	Not significant
Cities in Hurricane Florence	Education (%)	88.6	90.0	Not significant
	Unemployment (%)	5.3	5.6	Not significant
	Language (%)	6.0	4.9	Not significant
	Minority (%)	23.3	19.3	Not significant
	Medium income ($)	71,966	55,978	Not significant

Figure 4 characterizes the concentration of social media attention by showing the scaling laws followed by different types of tweets for areas in Hurricane Harvey (Figure 4a) and Florence (Figure 4b). Scaling laws in super-neighborhoods follow power-law relationships of the form $y \approx x^{\beta}$, where x is the population of a city or super-neighborhood, y is the number of tweets, and β is the scaling exponent (Table 5). This result indicates that social media attention is concentrated in highly populated areas. Such variation of social media attention is the consequence of the nonlinearity of user behaviors in online social networks (Balland et al., 2020). Specifically, in the case of Hurricane Harvey (Figure 4a), the number of original tweets mentioned in an area grows as the $\beta = 0.67$ power of the population. For retweets, the retweets granted to a super-neighborhood scale sub-linearly with population with an exponent of $\beta = 0.91$. Similarly, the replies category grows as the $\beta = 0.60$ power of the population in an area, quotes scale as the $\beta = 0.66$ power of population, and total tweets scale as the $\beta = 0.86$ power of population. Hence, the hypothesis **H2a** is accepted.

We repeated this exercise by studying the scaling laws followed by different types of tweets for cities, and found coefficients similar (but a little bit greater) to those for super-neighborhoods (Table 5). Thus, the variation of social media attention at the city scale is slightly greater than that at the super-neighborhood scale. Despite these small differences, generally, the coefficients for the relationships in city and super-neighborhood scales are close to each other. Hence, the quantitative relationships and corresponding findings are scalable for different disasters and areas. The hypothesis **H2b** is accepted. In addition, almost all coefficients are smaller than 1, which indicates a sub-linear relationship between population and social media attention in the logarithmic space. This implies that that the bias of social media attention regarding the size of the cities is weak (Barabási, 2013). The attention variation is greater among less populated areas than that among populous areas.

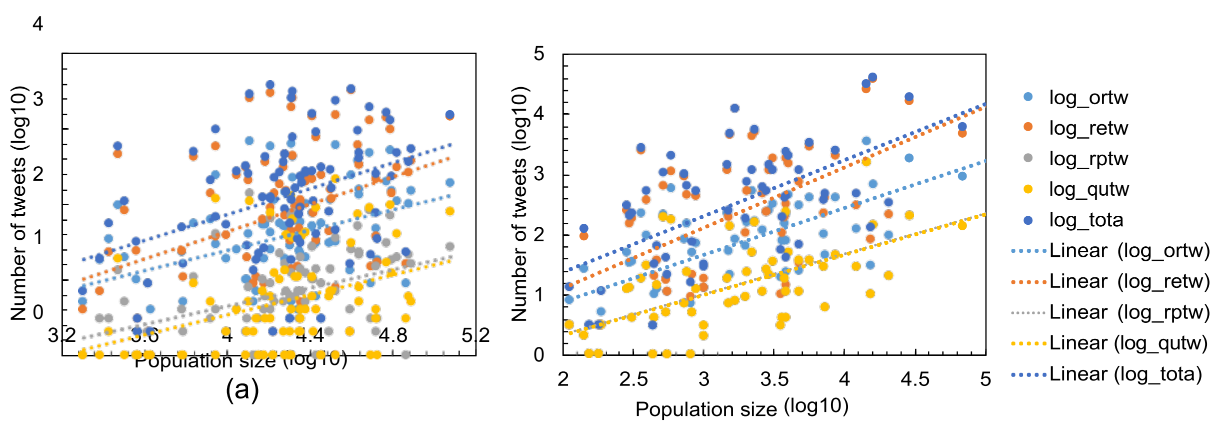


Figure 4. The relations between number of tweets and population for super-neighborhoods in Hurricane Harvey (a) and cities in Hurricane Florence (b). Here, "log_ortw" represents the logarithmic number of original tweets; "log_retw" represents the logarithmic number of retweets; "log_rptw" represents the logarithmic number of replies; "log_qutw" represents the logarithmic number of quotes; and "log_tota" represents the number of all types of tweets.

Table 5. The coefficients in the relations between population and different types oftweets.

Variables	Coefficient (β)	
	Super-neighborhoods	Cities
Original tweets	0.67***	0.78***
Retweets	0.91***	1.00***
Replies	0.60***	0.67***
Quotes	0.66***	0.65***
Total tweets	0.86***	0.93***

Note: * *P<0.05, ** P<0.01, *** P<0.001.*

Next, we applied the BERT-based machine learning model to label the tweets with humanitarian topics and investigated whether the concentration of social media attention has variations regarding information topics and areas.

Figure 5a shows the mean percentages of the tweets in each topic for groups of areas with different levels of attention. The figure reveals that, in general, high attention areas have slightly more proportions of informative tweets than those of low attention areas. However, this is not statistically significant. So, the average percentages of tweets in each topic are consistent in areas with different levels of attentions. The same pattern is observed at the city scale for Hurricane Florence (Figure 5b). By conducting pairwise comparisons for the mean percentages of tweets of different topics, we find that the tweets related to rescue and donation efforts are predominant (Table 6). The pattern is less significant at the city scale than that on the super-neighborhood scale during Hurricane Harvey. But, some of the findings are still scalable. For example, compared to the difference between infrastructure and rescue information, users paid more attention to rescue and donation related information. Besides, the information related to affected individuals and infrastructure damages account for similarly low percentages of the total amount of tweets, and do not have significant differences regarding the proportions of the tweets. The consequence of this is that social media pays more attention to posting and spreading rescue and donation information in for disaster-affected areas. Hence, the hypotheses **H3a** and **H3b** are both accepted.

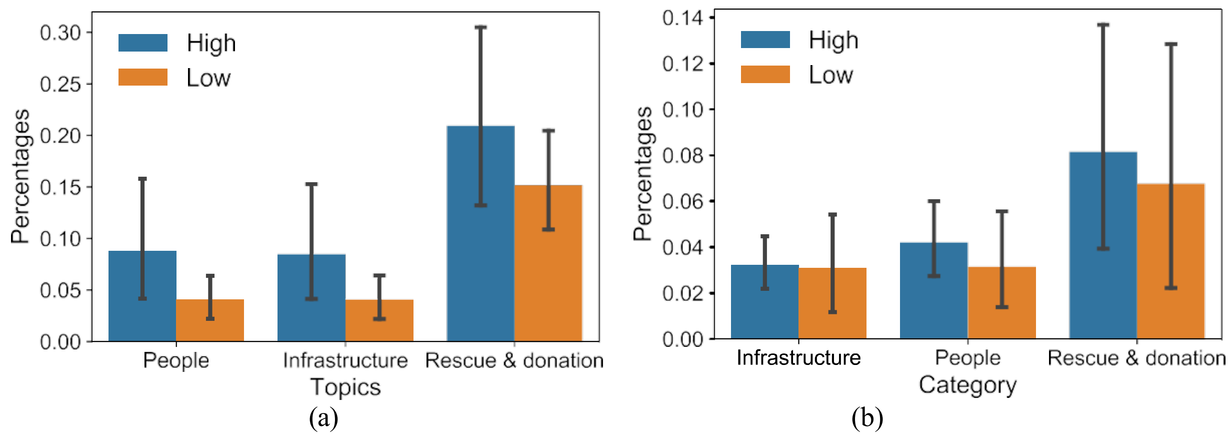


Figure 5. The percentages of the tweets with different topics for areas with different levels of social media attention: (a) super-neighborhoods in Hurricane Harvey; (b) cities in Hurricane Florence.

Table 6. The differences and significance of social media attention for different types of tweets.

Variables	*P-value*	
	Super-neighborhoods	Cities
Infrastructure vs. People	Not significant	Not significant
Infrastructure vs. Rescue & donation	***	*
People vs. Rescue & donation	***	Not significant

Note: ** P<0.05, ** P<0.01, *** P<0.001.*

CHAPTER VI

DISCUSSION AND CONCLUSION

This empirical study examined the influence of sociodemographic factors and damage claims on the attention of social media on areas in times of disasters. The core idea of this paper is that the population of an area explains the variations in the degree to which social media attention concentrates. We show this relationship to be true for all types of tweets and to be independent of other sociodemographic factors. We argue that social media users tend to pay more attention to populous areas. A sub-linear scaling law is adopted to quantify this relationship. The sub-linearity suggests that the variation of social media attention vary in sparsely-populated areas, while the attention is concentrated and constant among populous areas. For the topics of the tweet information, more attention was payed to the tweets related to rescue and donation efforts. Such variation is consistent in both super-neighborhood and city scales so that the quantified relationships can be applied to different scales.

The quantitative findings in this study have theoretical contributions to a more precise understanding of the bias of social media users in posting and spreading situational information in disasters. First, different from prior studies using geotagged tweets, this study is the first to measure social media attention by investigating the content of the tweets. The variation of social media attention across different areas confirms the existing studies using geotagged tweets, but the patterns are different. From the content-based analysis, the only related factor is the population size (the quantity of damage claims is also related to the quantify of tweets, but it is also related to the population size.). This finding contradicts existing studies that social media pay less attention to vulnerable populations (Vaughan & Tinker, 2009).

The theoretical findings can provide a guideline for other proxy studies of spatial analysis, which can explore the applications of social media data for disaster management. This study contributes to the scant literature concerning effectiveness of social media for needs and disruption assessment. For example, studies leverage social media to assess the disruptions and needs in disasters should take the population variation into consideration. Existing techniques and tools, such as event detection and topic modeling for understanding disruptions and needs, usually miss the representativeness issue of social media data. This issue would cause biases in these techniques (Du, Yang, Zou, & Hu, 2019). The evidence of the bias of social media attention present in this study is able to point out the source of the bias. By mitigating the spatial concentration of the data, existing machine learning models will be more effective to capture the situation from social media in disasters.

The idea that social media attention is concentrated more in populous areas poses a range of questions for planners and policymakers. It tells us that emergency management agencies and relief organizations need to rethink their response strategies. The effectiveness of the disaster management may well be dominated by those agencies that succeed at developing relief policies. Mega-areas are required to consume and process vast amounts of resources. This can be observed from the results for the variation of social media attention in information topics. The results in section 5 show that the mean percentages of the tweets in different topics remain constant across areas with different levels of social media attention. However, since the tweets are concentrated in populous areas, the variation of the absolute number of tweets for each topics is amplified among areas with various population size. This signifies that the rescue and donation related information concentrates in populous areas. This may lead to an important societal problem that the allocation of relief resources and the prioritization of response actions might be unfair due to the unequal attention paid by the public to different areas. For example, the public information posting and spreading on social media increases the exposure of disaster situations of populous areas, which would further attract more attention of the relief organizations. However, little would be known about the situation in sparsely-populated areas. Although the damages and needs in populous areas might be more severe than those of sparsely populated areas, the variation of social media attention may exacerbate the inequality of the disaster treatment in terms of the efficiency of response and the sufficiency of supplies. Hence, if the social media attention and the population size cannot be dissociated, the spatial inequality observed among populous and sparsely-populated areas is likely to increase in real-word disaster management. Policymakers and response managers must recognize such inequality both within cities and between cities.

Due to the inequality of social media attention, we also face critical questions regarding the future of mitigation planning and the distribution of infrastructure facilities to cope with disasters for sparsely-populated areas. Since the situation in sparsely-populated areas are less exposed, the capabilities of damage resistance and disaster resilience for these areas should be taken into consideration for mitigation planning. With less external support for disaster response, people living in sparsely-populated areas should be given equitable access to the relief resources made available to populous areas, which requires a high accessibility to transportation systems during disasters. This suggests that the infrastructure from sparsely-populated areas to populous areas needs to have a greater capacity to absorb the negative impacts of disasters, such as potential congestion due to evacuation and floodwaters. In addition, people in populous area have greater accessibility to situational information due to high volume of social media attention, while people in sparsely-populated areas may not be able to capture a complete picture of the situation without sufficient social media posts. Therefore, facilities that can enable the cohesion of people in sparsely-populated to enhance their situation awareness is important. Planners need to consider these important insights for improving the disaster resilience and inequality of different areas.

Like any social media research, our study has limitations that need to be taken into consideration. The descriptive nature of our analysis only focuses on examining the relationships but does not provide a clear indication of the causes leading to increases in the spatial concentration of social media attention in disasters. The topics in the texts of the tweets might provide cues for understanding the causes of this concentration pattern. For example, rescue and donation efforts might be more concentrated in populous areas, which could lead to a concentration of social media attention. However, the concentration of other relevant tweets in populous areas cannot be clearly explained. Hence, future studies can look into the causal effects of societal characteristics on social media attention in disaster-affected areas. Second, while this study obtains the quantitative findings in both super-neighborhood and city scales for two national-wide representative disasters, the next question is how much these patterns vary from these cases to other cases. Through adopting the analysis in other disaster events, future studies can evaluate the universality of these patterns regarding social media attention from our study.

REFERENCES

Alam, F., Ofli, F., & Imran, M. (2018). CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. Retrieved from http://arxiv.org/abs/1805.00713

American FactFinder. (2020). American FactFinder. Retrieved February 19, 2020, from https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Balland, P. A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., & Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0803-3

Barabási, A.-L. (2013). Network science. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*. https://doi.org/10.1098/rsta.2012.0375

Caragea, C., Silvescu, A., & Tapia, A. H. (2016). Identifying informative messages in disaster events using Convolutional Neural Networks. In *Proceedings of the International ISCRAM Conference*.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You Are Where You Tweet : A Content-Based Approach to Geo-locating Twitter Users. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 759–768. https://doi.org/10.1145/1871437.1871535

Chou, W. Y. S., Hunt, Y. M., Beckjord, E. B., Moser, R. P., & Hesse, B. W. (2009). Social media use in the United States: Implications for health communication. *Journal of Medical Internet Research*. https://doi.org/10.2196/jmir.1249

Cutter, S. L., Boruff, B. J., & Shirley, W. L. (2003). Social vulnerability to environmental hazards. *Social Science Quarterly*. https://doi.org/10.1111/1540-6237.8402002

Du, M., Yang, F., Zou, N., & Hu, X. (2019). Fairness in Deep Learning: A Computational Perspective. Retrieved from http://arxiv.org/abs/1908.08843

Fan, C, Wu, F., & 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. https://doi.org/10.1109/ACCESS.2020.2965550

Fan, Chao, Jiang, Y., & Mostafavi, A. (2020). Social sensing in disaster city digital twin: an integrated textual-visual-geo framework for situational awareness during built environment disruptions. *Journal of Management in Engineering*, 1–12.

Fan, Chao, & Mostafavi, A. (2019). A graph-based method for social sensing of infrastructure disruptions in disasters. *Computer-Aided Civil and Infrastructure Engineering*, *34*(12), 1055–1070. https://doi.org/10.1111/mice.12457

Fan, Chao, Mostafavi, A., Gupta, A., & Zhang, C. (2018). A System Analytics Framework for Detecting Infrastructure-Related Topics in Disasters Using Social Sensing. In I. F. C. Smith & B. Domer (Eds.), *Advanced Computing Strategies for Engineering* (pp. 74–91). Cham: Springer International Publishing.

Fischer-Preßler, D., Schwemmer, C., & Fischbach, K. (2019). Collective sense-making in times of crisis: Connecting terror management theory with Twitter user reactions to the Berlin terrorist attack. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2019.05.012

Frigerio, I., & De Amicis, M. (2016). Mapping social vulnerability to natural hazards in Italy: A suitable tool for risk mitigation strategies. *Environmental Science and Policy*. https://doi.org/10.1016/j.envsci.2016.06.001

Ghani, N. A., Hamid, S., Targio Hashem, I. A., & Ahmed, E. (2018). Social media big data analytics: A survey. *Computers in Human Behavior*, (July). https://doi.org/10.1016/j.chb.2018.08.039

Hecht, B., & Stephens, M. (2014). A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.

Hidayatullah, A. F., Aditya, S. K., Karimah, & Gardini, S. T. (2019). Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA). In *IOP Conference Series: Materials Science and Engineering*. https://doi.org/10.1088/1757-899X/482/1/012033

Horsburgh, J. S., Morsy, M. M., Castronova, A. M., Goodall, J. L., Gan, T., Yi, H., … Tarboton, D. G. (2016). HydroShare: Sharing Diverse Environmental Data Types and Models as Social Objects with Application to the Hydrology Domain. *Journal of the American Water Resources Association*. https://doi.org/10.1111/1752-1688.12363

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Eighth International AAAI Conference on Weblogs and Social Media*. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2014). *Processing Social Media Messages in Mass Emergency: A Survey*. https://doi.org/10.1145/2771588

Jiang, Y., Li, Z., & Ye, X. (2019). Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartography and Geographic Information Science*, *46*(3), 228–242. https://doi.org/10.1080/15230406.2018.1434834

Keating, R. T., Hendy, H. M., & Can, S. H. (2016). Demographic and psychosocial variables associated with good and bad perceptions of social media use. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2015.12.002

Kibanov, M., Stumme, G., Amin, I., & Lee, J. G. (2017). Mining social media to inform peatland fire and haze disaster management. *Social Network Analysis and Mining*. https://doi.org/10.1007/s13278-017-0446-1

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Hentenryck, P. Van, Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advance*, (March), 1–12. https://doi.org/10.1126/sciadv.1500779

Kumar, S., Hu, X., & Liu, H. (2014). A behavior analytics approach to identifying tweets from crisis regions. *Proceedings of the 25th ACM Conference on Hypertext and Social Media - HT '14*, 255–260. https://doi.org/10.1145/2631775.2631814

Li, H., Caragea, D., Caragea, C., & Herndon, N. (2018). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*. https://doi.org/10.1111/1468-5973.12194

Li, Z., Wang, C., Emrich, C. T., & Guo, D. (2018). A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods. *Cartography and Geographic Information Science*, *45*(2), 97–110. https://doi.org/10.1080/15230406.2016.1271356

MacEachren, a. M., Robinson, a. C., Jaiswal, a., Pezanowski, S., Savelyev, a., Blanford, J., & Mitra, P. (2010). Geo-Twitter Analytics: Applications in Crisis Management. *Proceedings of the 25th International Cartographic Conference*.

Madianou, M. (2015). Digital Inequality and Second-Order Disasters: Social Media in the Typhoon Haiyan Recovery. *Social Media and Society*. https://doi.org/10.1177/2056305115603386

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population Bias in Geotagged Tweets. *ICWSM-15 Workshop: Standards and Practices in Large-Scale Social Media Research*, 18–27. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662

Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-Time Crisis Mapping of Natural Disasters Using Social Media. *IEEE Intelligent Systems*, *29*(2), 9–17. https://doi.org/10.1109/MIS.2013.126

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from twitter's streaming API with Twitter's firehose. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*.

Neubaum, G., Rösner, L., Rosenthal-Von Der Pütten, A. M., & Krämer, N. C. (2014). Psychosocial functions of social media usage in a disaster situation: A multi-methodological approach. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2014.01.021

Paul, S., Ghebreyesus, D., & Sharif, H. O. (2019). Brief Communication : Analysis of the Fatalities and Socio-Economic Impacts Caused by Hurricane Florence, *2005*, 1–12. https://doi.org/10.3390/geosciences9020058

Pulcinella, J. A., Winguth, A. M. E., Allen, D. J., & Dasa Gangadhar, N. (2019). Analysis of Flood Vulnerability and Transit Availability with a Changing Climate in Harris County, Texas. *Transportation Research Record*. https://doi.org/10.1177/0361198119839346

Ragini, J. R., Anand, P. M. R., & Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, *42*, 13–24. https://doi.org/10.1016/j.ijinfomgt.2018.05.004

Rahimi, A., Cohn, T., & Baldwin, T. (2015). Twitter user geolocation using a unified text and network prediction model. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.3115/v1/p15-2104

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*. https://doi.org/10.1080/15230406.2017.1356242

Seppänen, H., & Virrantaus, K. (2015). Shared situational awareness and information quality in disaster management. *Safety Science*, *77*, 112–122. https://doi.org/10.1016/j.ssci.2015.03.018

Sloan, L., & Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PLoS ONE*, *10*(11), 1–15. https://doi.org/10.1371/journal.pone.0142209

Sutton, J., Gibson, C. Ben, Phillips, N. E., Spiro, E. S., League, C., Johnson, B., … Butts, C. T. (2015). A cross-hazard analysis of terse message retransmission on Twitter. *Proceedings of the National Academy of Sciences*, *112*(48), 14793–14798. https://doi.org/10.1073/pnas.1508916112

Vaughan, E., & Tinker, T. (2009). Effective health risk communication about pandemic influenza for vulnerable populations. *American Journal of Public Health*. https://doi.org/10.2105/AJPH.2009.162537

Wang, Q. (2015). Human Mobility Perturbation and Resilience in Natural Disasters. Retrieved from https://vtechworks.lib.vt.edu/handle/10919/51955

Xiao, Y., Huang, Q., & Wu, K. (2015). Understanding social media data for disaster management. *Natural Hazards*, *79*(3), 1663–1679. https://doi.org/10.1007/s11069-015-1918-0

Yang, Y., Zhang, C., Fan, C., Yao, W., Huang, R., & Mostafavi, A. (2019). Exploring the emergence of influential users on social media during natural disasters. *International Journal of Disaster Risk Reduction*, *38*, 101204. https://doi.org/10.1016/J.IJDRR.2019.101204

Zhang, C., Fan, C., Yao, W., Hu, X., & Mostafavi, A. (2019). Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2019.04.004

Zou, L., Lam, N. S. N., Shams, S., Cai, H., Meyer, M. A., Yang, S., … Reams, M. A. (2018). Social and geographical disparities in Twitter use during Hurricane Harvey. *International Journal of Digital Earth*, *0*(0), 1–19. https://doi.org/10.1080/17538947.2018.1545