2020

University of North Carolina Wilmington

Master of Science in

Computer Science and Information Systems

Proceedings

https://csbapp.uncw.edu/mscsis

TEXT ANALYTICS AND SPATIAL VISUALIZATION OF SOCIAL MEDIA DATA DURING A DISASTER LIFECYCLE: THE CASE OF HURRICANE DORIAN

Cyrus Goudarzi

A Thesis Submitted to the
University of North Carolina Wilmington in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

Department of Computer Science
Congdon School of Supply Chain, Business Analytics, and Information Systems

University of North Carolina Wilmington

2020

Approved by

Advisory Committee

Jeffrey Cummings _____

Sudip Mittal _____

Minoo Modaresnezhad
_____
Chair

Accepted By

_____
Dean, Graduate School

TABLE OF CONTENTS

ABSTRACT

Social media platforms have become an increasingly prevalent means of human communication ever since they grew in popularity across the world. As a result, data scientists have more text data available than ever before. This abundance of data is advancing the field of text data analysis and natural language processing by allowing data scientists to evaluate larger data sets within a specific context. One specific context that has not been as explored is the evaluation of text data generated during a natural disaster. This study's contributions to previous studies within the disaster context include an exploration of metadata extraction of Twitter user data, an evaluation of current geographic information systems technologies, and a dashboard design for monitoring Twitter activity throughout the disaster lifecycle. This dashboard concept uses visualizations of text data analysis techniques in contrast to each other for the overall purpose of disaster monitoring.

The focus of this project is to analyze a dataset of tweets collected before and during Hurricane Dorian using the hashtag "#HurricaneDorian." The goal of this research is to provide a dashboard concept for disaster management through text mining and spatial analysis techniques using Twitter data. This research also seeks to provide a thorough analysis and summary of text data collected surrounding Hurricane Dorian. Specifically, this research aims to answer the following research questions: *H*ow can social media (Twitter) be utilized to identify, process, and comprehend critical elements of an incident or situation during a natural disaster? *H*ow does the analysis of geotagged data compare and contrast to the analysis of non-geotagged data within the disaster context?

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1: INTRODUCTION

Quantitative text data analysis is a growing field of research due to the massive amount of text data available from social media platforms. These platforms are actively changing the way that individuals, institutions, and organizations alike are experiencing and understandings temporal events. Since different social media platforms are actively providing collections of text data from users, researchers can now apply natural language processing and text data analysis techniques to a broader variety of situational contexts than in the past. A natural disaster is one of these categories of events that can be better understood through text data analysis and visualization. This study continues previous studies of text data analysis and visualization by applying these techniques on Twitter data from Hurricane Dorian.

In the fall of 2019, Hurricane Dorian ravaged the Caribbean and Florida, resulting in disastrous situations for all the individuals in that area. The Hurricane then proceeded its way up the east coast, ultimately affecting southeastern U.S. states of Georgia, South Carolina, North Carolina, and Virginia. Fifteen years ago, we would not have had large sets of text documents surrounding an event; however, today, platforms like Twitter actively record the experiences of locals (victims) affected by a disaster. In this case, Twitter provided a platform for communication among individuals affected by Hurricane Dorian, who had access to these technologies.

By applying a text data analysis and spatial-temporal visualization, we can hopefully understand new trends regarding the communities affected by the Hurricane and provide insight for further preparation for natural disaster relief agencies. Previous studies have shown how existing sentiment classification models have strengths and weaknesses when being applied in a disaster context. One study, for example, found that members of the disaster relief team believed sentiment analysis of Twitter data would help in understanding locals (victims) situational

awareness; however, their current collection of data was too generalized to be useful. As a result, the researchers added layers of classification to filter out data that was irrelevant to understanding victims' situational awareness (Utz, Schultz, & Glocka, 2013). The results of such studies could be used in prioritizing tweets that are closely related to emergency management than traditional classification models. An advanced model could influence and improve disaster management decision-making and operations. This study exemplifies how to use text data analytics techniques like sentiment analysis and topic modeling to optimize the flow of information provided for emergency management teams.

In order to continue to push forward the contextualization of text analysis in disaster events, this research project seeks to clean and process the given collection of text objects, compare and contrast results between geotagged data and the entire dataset, and design a dashboard to display a collection of different text analysis visualizations regarding the event. This experiment will use Python (python library Pandas) for importing and parsing JSON data, as well as for pre-processing and cleaning the given data. The application will then use the stanfordNLP library to implement sentiment analysis and topic modeling techniques as well as Tableau in order to visualize the text data analysis results between differing datasets.

The dashboard's goal is to provide a visual and graphical representation of the Twitter activity throughout Hurricane Dorian so that future trends and insights of disasters can be utilized to improve disaster relief efforts. The dashboard will visually represent three separate types of text analysis, including sentiment analysis, topic modeling, and word frequency in order to help us learn more about trends of social media throughout disasters. For this research, the tweets will be concentrated from the states of Virginia, North Carolina, South Carolina, Georgia, and Florida. The dashboard design for this project is developed in hopes of improving future

disaster relief efforts by providing a user interface to highlight quantitative and qualitative insights of Twitter activity throughout a disaster, like Hurricane Dorian.

CHAPTER 2: REVIEW OF LITERATURE

Microblogging services have been a convenient platform for victims of natural disasters seeing that many victims of a natural disaster may have access to cellular support and a smartphone. Traditionally, communication throughout a natural disaster flows in one direction from authorities to victims; however, social media platforms like Twitter have opened bilateral paths of communication, enabling victims to initiate conversations. Access to social media has increased the ability of victims to warn others of dangerous situations in their region or to raise awareness and funds towards damages of the natural disaster. Hurricane Katrina elaborated on this concept and applied it to the earthquakes of Haiti in 2010. While Katrina victims did not have the same access to social media platforms in 2005, the authors argued that overall greater awareness of the damages resulted in a more successful fundraising campaign by the Red Cross. Not only could witnesses of the earthquake share crisis in their region, but they also were able to raise greater global awareness of the earthquake's victims through these web platforms (Beigi, Hu, Maciejewski, & Liu, 2016).

In contrast to Hurricane Katrina, for example, Hurricane Sandy was a natural disaster occurring in 2012 that provided an abundance of text data available for data scientists across the world. Hurricane Sandy happened during the 2012 Atlantic hurricane season and ultimately affected 24 U.S. States causing an estimated $70.2 billion in damages (CNN, 2013, Para. 1). In the aftermath of Hurricane Sandy, many information scientists took advantage of the available data from various social media sources to learn more about topics and trends throughout the occurrence of the event.

Data scientists are also using text data to compare how different types of information travel throughout an event. In order to look for early warning signs of a disaster, one study

prioritized *awareness* as a category of information traveling between individuals. The group implemented a sensor method looking at attributes of information for each user such as entry time, total number of messages, and counts of friends and followers. "For the sensor method to work, the relationship must exist between users' entry times and their topological (node degree) or behavioral (activity) characteristics." This social network was able to identify distributions of users with higher activity and numbers of followers. In looking at awareness, the study also found that "the sensor method results in the awareness advantage on a scale between 3 and 26 hours, depending on the sample size and geographical origin of the groups" (Kryvasheyeu, Chen, Moro, Van Hentenryck, & Cebrian, 2015). This analysis describes how *awareness* was measured throughout the spread of information regarding the disaster. This kind of analysis was previously unavailable without Twitter attributes like geotagging, timestamps, and attributes of user accounts.

Some studies categorize different approaches to evaluate individual awareness throughout a natural disaster by dividing the social media data between the following categories of (a) situational awareness: identifying, processing, and comprehending critical elements of an incident or situation for insight, and (b) information sharing: showing how people behave and share information in social media regarding the disasters (Beigi et al., 2016). Generally, social media data could potentially optimize the supply-chain used for getting resources to the public.

While the social media platform is used to spread information throughout a natural disaster, it often results in misinformation between users. Previous studies have shown how social media just as easily has spread false information that can influence a victims' decision making process. One study that sought out to find the influence of content ambiguity on the spread of rumors within social media by using logistic regression analysis. The study concluded

in their findings that "while content ambiguity does not contribute to rumormongering, source ambiguity does so very significantly…messages in the category of source ambiguity frequently resembled third-person situation reports without sources being attached" (Oh, Agrawal, & Rao, 2013). Misinformation is just as influential of a problem to victims' safety as a lack of information; therefore, improving understanding of the flow of information for victims can further help disaster relief agencies to reduce these rumors.

For example, in Hurricane Irma of fall 2017, a rumor circulated throughout Twitter that there is a benefit to shooting guns into the Hurricane in order to reduce stress and boredom. While this could be taken as a facetious statement, it ultimately gained enough circulation where the Sherriff department of Pasco County had to issue an illustration of the dangers of firing bullets into the hurricane to help prevent this behavior. (Pasco County Sherrif on Twitter, 2017) Another example includes a recurring rumor that occurred both in 2017's Hurricane Irma and 2018's Hurricane Florence that at some point, sharks were lifted into the hurricanes themselves, providing additional threat and hysteria to users. The graphic used in these rumors were directly from a sci-fi move series *Sharknado*; however, some users did believe these rumors from learning about them through Twitter. (Snopes, Para. 2, 2017). Although outside of the scope of the current study, we believe studying misinformation has an important role when we are studying natural disaster social media data.

CHAPTER 3: METHODOLOGY

Twitter Data

While APIs or Application Programming Interfaces have become a crucial piece of successful web application business models they have also influenced the research world due to their abundance of data. Twitter's API, for example, has large datasets of text data available for developers to process, clean, and analyze how they wish. This availability is highly influential in the world of text data analysis since Twitter's API provides a variety of data sources through its international audience.

Twitter has many APIs for developers to choose from and breaks them down into different categories, such as Standard APIs, Premium APIs, Enterprise APIs, and Ads APIs. An example of their Standard API would be the Account Activity API that allows developers to integrate Twitter data regarding activities like tweets by a user, replies by users, blocks by a user, and other account-related activities of a user. Twitter's API has a variety of features regarding interacting with the Twitter platform for developers, including but not limited to "Tweets, Media, Direct Messages, Trends, Geo, Ads" and more (Twitter). Having a variety of API features certainly helps the financial benefactors of this API's implementation; however, they also help developers pursuing research in text data analysis.

The format of this API is in a JSON text format that allows the developers to manipulate the JSON data as they like. "JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages" (JSON, 2019, Para. 1). It is important to recognize that having a text format could be suggested as a reason for JSON's popularity within the world of Web APIs. JSON is built upon two popular data structures consisting of an (a) object: or collection of name/value pairs, (b) array: an ordered list of values. The organization behind JSON writes on their homepage, "these are universal data structures.

Virtually all modern programming languages support them in one form or another. It makes sense that a data format that is interchangeable with programming languages also be based on these structures." (JSON 2019, Para. 3) On the other hand, perhaps its JSON's architecture to credit for its popularity; regardless, Twitter's API breaks down individual tweets as JSON objects when parsing and delivering requests.

The data used in this study was collected through Twitter's API from August 24, 2019 to September 10, 2019 using the hashtag #HurricaneDorian. While the text data provided by Twitter provides a variety of JSON objects containing information regarding the tweet, the following categories will be considered for the purpose of this research project:

Twitter JSON Objects	
"text"	280 character text message of the tweet
"coordinates"	longitudinal, latitudinal coordinates of the tweet from browser or device
"place"	Specific, named locations with corresponding geo coordinates
"timestamp_ms"	Timestamp of when the tweet was posted
"reply_count"	The number of replies to a given tweet
"created_at"	Timestamp of when the tweet was published
"friends_count"	The number of friends belonging to author of tweet
"verified"	Whether the twitter account is a verified account or not
"geo"	A set of geocoordinates of the location of the user when the tweet was published
"favorites_count"	The number of favorites assigned to tweet by other users
"location"	A location's name if user checks-in to pre-existing location for

	tweet
“retweeted”	Whether the tweet has been retweed by the authenticating user
“description”	Written description by user, about user profile
“geo_enabled”	Boolean whether geotagging services are allowed for user’s tweets

Table 1. Twitter JSON Objects

The “text” object will provide the text messages for natural language processing. In contrast, the “coordinates” category will provide the latitudinal and longitudinal necessary for the spatial visualization portion of the project. Lastly, the “timestamp_ms” category will be used to correctly orchestrate the order of tweet submissions throughout the temporal visualization portion of the analysis.

For the current research, the python library Pandas will be used in order to parse the JSON twitter objects, pre-process/clean the data, and to organize the data into data frames. Below is an example of an individual tweet at random from the sample in which a script using Python library Pandas has created a data frame of the individual attributes of the given tweet.

```
In [18]: runfile('C:/Users/Cyrus Goudarzi/TwitterCleanTutorial/Script.py', wdir='C:/Users/Cyrus
Goudarzi/TwitterCleanTutorial')
<--------------USER OBJECT----------------->
{'id': 292698490, 'id_str': '292698490', 'name': 'IAMSPonline', 'screen_name': 'IAMSPonline',
'location': 'Global', 'url': 'http://www.iamsponline.org', 'description': 'The International
Association of Maritime Security Professionals (IAMSP) is a global non-profit association for MARSEC
professionals. RT ≠ Endorsement. #marsec', 'translator_type': 'none', 'protected': False, 'verified':
False, 'followers_count': 3401, 'friends_count': 2966, 'listed_count': 135, 'favourites_count': 8726,
'statuses_count': 21828, 'created_at': 'Wed May 04 02:29:59 +0000 2011', 'utc_offset': None,
'time_zone': None, 'geo_enabled': False, 'lang': None, 'contributors_enabled': False, 'is_translator':
False, 'profile_background_color': 'C0DEED', 'profile_background_image_url': 'http://abs.twimg.com/
images/themes/theme1/bg.png', 'profile_background_image_url_https': 'https://abs.twimg.com/images/
themes/theme1/bg.png', 'profile_background_tile': False, 'profile_link_color': '1DA1F2',
'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_color': 'DDEEF6', 'profile_text_color':
'333333', 'profile_use_background_image': True, 'profile_image_url': 'http://pbs.twimg.com/
profile_images/2840390402/7e023f83bc87bebf784e0991169948d5_normal.jpeg', 'profile_image_url_https':
'https://pbs.twimg.com/profile_images/2840390402/7e023f83bc87bebf784e0991169948d5_normal.jpeg',
'profile_banner_url': 'https://pbs.twimg.com/profile_banners/292698490/1356109953', 'default_profile':
True, 'default_profile_image': False, 'following': None, 'follow_request_sent': None, 'notifications':
None}
<--------------TWEET TEXT----------------->
RT @Go
```

Figure 1. Pre-processed Data Example- User Object and Tweet Text

Printed to the console are the tweet's message as well as the user object of the associated tweet.

Metadata and demographic extraction and analysis

Twitter User Characteristics

Some studies argue that one way to improve the accuracy of text data analytics from social media is to understand further and classify the information about the user. In this study, we will consider users' metadata as well as the text data. This emphasis on user metadata also can help identify the way in which victims of a given disaster are prioritizing the information they are receiving and sharing.

Understanding social network relationships between users has been shown to improve user-level sentiment analysis in opinion mining. Research has shown that Twitter users engage with others can influence the overall sentiment of a collection of text data. Including information about the user's relationship with other accounts has been shown to improve the overall sentiment classification as long as there is a strong correlation between user connectedness and shared sentiment" (Tan et al., 2011).

The relationships between Twitter users' metadata and their accounts can help disaster relief teams differentiate how victims are assessing their surroundings. This type of information could benefit the way that media outlets are used to disperse information from disaster relief agencies as well as improve the effectiveness of communication from victims to relief members.

A quantitative approach to evaluating user accounts within a social network is to look at the overall number of reactions involved among stakeholders. This evaluation of user activity was emphasized in a study regarding crisis communication throughout the Fukushima nuclear disaster where the authors wrote, "Due to the inherent conversational and transparent character of the social media tools, organizations using social media can deliver real time information to concerned stakeholders and can thereby alleviate the stress of the unknown. In this way, organizations are

meeting stakeholders' demands for timely and accurate information." (Utz et al., 2013). By evaluating a quantitative analysis on user account behavior we can ultimately see if there is a trend between a number of actions between accounts, the type of account (verified or non-verified), and its overall relationship to the trending sentiment or topic.

Text Filtering and Preprocessing

Another obstacle to accuracy in quantitative text data analytics is the increased amount of characters used in Tweeting. Having a surplus of characters like "Emojis" complicate our ability to classify text data since each symbol influences the overall sentiment. One study explains the different ways in which emojis can influence sentiment classification after finding that "considering Emoji in sentiment analysis help improve overall sentiment scores" (Ayvaz & Shiha, 2017). Deciding how to pre-process these unique characters is one of the biggest challenges in maintaining the quality of a text dataset. Other unique characters can include links to third-party websites and other abbreviations. "Twitter's 140 character limit has been cited as an aspect of the platform that can affect a lexicon's accuracy in sentiment classification" (Hu, Tang, Tang, & Liu, 2013). In a changing landscape of conversation, these nuances in language through social media ultimately can influence the quality in quantitative text data analytics and the efficiency of modern-day lexicons. It is important to establish these levels of classification within the pre-processing phase so that the text data quality is as unique as possible during the analysis.

Sentiment Analysis

In sentiment analysis, one is analyzing the overall feelings or opinions found using language within a message. Opinion mining is a similar concept where a researcher is actively seeking and measuring the presence of opinions within a specific data set. Sentiment analysis gained popularity after 2001, which could be deduced to a multitude of factors; however, some could include a rise in digital microblogging services, an abundance of accessible data sets

present from users throughout the internet, and a rise in machine learning abilities for natural language processing; “Much of the subsequent research self-identified as opinion mining fits this description in its emphasis on extracting and analyzing judgments on various aspects of given items. However, the term has recently also been interpreted more broadly to include many different types of analysis of evaluative text” (Pang & Lee, 2008). In 2019, sentiment analysis is commonly sought out information for organizations including measuring public opinion when it comes to market research for political, commercial, and academic organizations.

The most basic approach to sentiment from text data includes three categories of positive, negative, and neutral sentiments (Go, Bhayani, & Huang, 2009). Sentiment analysis has its own set of nuance; whereas researchers have worked on ways to improve the efficiency of terms such as subjectivity classification and polarity classification. While the most commonly used tools include polarity, strength, and emotion; opinion mining results are dependent on the data processing model that breaks down a given set of text data. In result, it can be suggested that potential threats to the integrity of sentiment analysis’s nuance include polarity reducing the dimensions of opinions as well as the conflicts that come from attempting to classify the polarity of a mixed emotion. (Bravo-Marquez, Mendoza, & Poblete, 2014)

A lexicon is one popular tool for categorizing the polarity of words in that it provides a collection of categories and their subordinate words that allows for the evaluation of each word within the given text dataset. It is also a popular pursuit of text data analysis research to create lexicons and implement them within sentiment analysis. While that was not a part of this study, I believe that studies regarding this topic can be helpful when wanting to understand issues disrupting the accuracy of sentiment analysis and areas in which this particular field of text data analysis can improve. For example, the study Building Lexicon for Sentiment Analysis from

Massive Collection of HTML Documents found in their pursuit of lexicon building that the majority of the errors in their polarity classification centered on neutral phrases. They also noted that the efficiency of their implementation of the given lexicon suffered from the lexicon's small size (Kaji & Kitsuregawa, 2007). While a small lexicon could be expected to perform worse than one larger, the larger concept reinforced by this study is that polarity classification becomes increasingly difficult among mixed emotions and characters.

Sentiment analysis and opinion mining in the context of natural disasters is a field that can be looked at from multiple views such as the operations of communications within a disaster zone or the sentiments conveyed to users outside of a disaster zone. Either way it is clear that natural disasters can affect the sentiments of individuals in each phase of the event whether preparation, survival, or cleanup. In *Performance of Social Network Sensors during Hurricane Sandy*, the authors further evaluated the sensor method through matching each word of a message against a pre-existing dictionary and applying specific weights of emotion. The authors describe their sensor method as follows: "Total weights are calculated, normalized by the word count and returned as either a relative sentiment (average of all scores taking into account their sign) or an absolute sentiment (average of absolute values)" (Kryvasheyeu et al., 2015). By applying this sensor method in parallel with a sentiment analysis, the authors found that the sensor method groups they implemented did perform more accurately than the control groups of their experiment consisting of randomly generated datasets. These sensor groups can be further applied as a methodology in tracking flow of information between disaster participants regarding emotional and situational awareness. The study also found that "factors like a relatively short time scale and strong geographical nature of a disaster affect performance of the method" (Kryvasheyeu et al., 2015).

Topic Modeling Techniques

While there are many various topic modeling techniques in the world of text analysis, this research paper will only be focusing on models that can be defined as a "generative probabilistic model for collections discrete data such as text corpora" (Blei, Ng, & Jordan, 2003). From a high-level view, this family of models generates probability distributions across a body of text data through different variable implementations. For example, a researcher may implement a variable like *number of topics* which could be applied across multiple levels of a collection of documents in order to find different probability distributions between different layers of the data set. The topic modeling techniques focused within this study will have at least three dimensions including documents, topics, and words. In this experiment, we define topics and documents as, "…a mixture over words where each word has a probability of belonging to a topic. And a document is a mixture over topics, meaning that a single document can be composed of multiple topics" (Roberts, Stewart, & Tingley, 2019).

Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation (LDA) can be described as a "three-level hierarchical Bayesian model" referencing the way that LDA uses a hierarchy of documents, topics, and words in order to construct a statistical model. In the latent Dirichlet allocation method, the most outer view of the text collection consists at the document level. In the document level, documents are constructed out of probability distributions of a set number of topics. This input of the set number of topics is also known as K. This variable K is considered to be the dimensionality of the topic variable z because it is the set number of attributes or topics that occur per document object. Some argue that the required input of a set number of topics is one of the weaknesses of the LDA method

because the user consistently alters the distribution of probabilities occurring at the document level.

The next inner level within the statistical hierarchy is at the topic level. Each topic level is constructed out of a probability distribution of words. The latent Dirichlet allocation method applies a set number of topics K onto a corpus of words in order to construct a topical level within each document. By organizing the individual words into groups within the document, the LDA method has broken down the collection into three tiers where each word belongs to a topic and each topic belongs to a document.

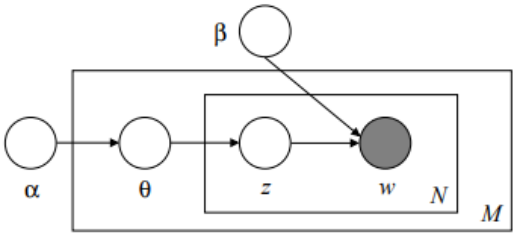


Figure 2. Graphical Model representation of LDA

One commonly referenced notation for latent Dirichlet allocation is known as plate notation. Plate notation strives to visually illustrate the model parameters and their dependencies. In the figure below the rectangle M represents the total number of documents within a corpus while the smaller rectangle N represents a set of words within a document. Plate notation also helps illustrate the three different levels to LDA modeling by graphically showing that alpha and beta are sampled once at the corpus or document level. The topical level of latent Dirichlet allocation is what separates LDA from the traditional topic clustering model which only would utilize two-levels.

LDA is only one of the multitude of statistical models that are commonly used for text data analytics. Others models statistical models differ from LDA generally in the number of variables included. For example, a model such as Structural Topic Modeling (STM) build upon

the three-level hierarchy found within LDA by including a fourth level of document metadata, which is additional information about each document. Structural Topic Modeling's ultimate goal is to estimate a relationship between topics and document metadata; however, that is the biggest difference between Latent Dirichlet allocation and structural topic modeling. STM indexes each document within the collection with document metadata in order to provide additional insights on the dataset. These additional relationships are known as covariates and can be assed as follows: "metadata that explain topical prevalence are referred to as topical prevalence covariates, and variables that explain topical content are referred to as topical content covariates." (Roberts et al., 2019)

GIS

The growing field of Geographic Information Systems (GIS) is changing the way that institutions and individuals understand and experience natural disasters across the world. The application of information systems tools within the geology field has given data scientists further opportunities disaster risk assessment data acquisition, data management and processing, spatial analysis, data modeling and simulation and is increasingly becoming an important tool in disaster monitoring. This is especially important for the natural disaster context as spatial visualization of topics within text data can provide valuable insight for improving the relief efforts of a given disaster. GIS Mapping tools like ArcGIS have enabled developers the ability to visualize variables regarding communities throughout a given crisis.

GIS is significant for the disaster context as previous studies have been able to improve disaster preparation and relief efforts with GIS tools. In some cases, scientists have been able to create maps that include spatial visualizations of risk assessment probabilities for a given region. (Wang, Tu, Liu, & Zhao, 2018) Another study looked at geotagged Twitter data in an effort to understand human travel throughout the disasters Hurricane Sandy, Typhoon Wipha, and

Typhoon Haiyan in Tacloban. By developing a geographic boundary for each physical event of the disaster, the researchers were able to filter geocoordinates of tweets occurring within each boundary which tweets belonged to victims of the event. The study concluded "that tropical cyclones change the frequency of human travels, often increasing short distance trips and suppressing long distance ones" (Qi, W., & John E., T., 2014). Another study also found that a geographic context on non-geotagged data can also be effective in understanding physical circumstances of victims within a disaster. The study developed "a multi-elemental location inference method" for searching for traces of geographic keywords through non-geotagged, Twitter datasets. While the study found "[the method] was able to successfully infer the location of 87% of the sample tweets…. which is a significant improvement compared with that of the current methods that can predict the location with much larger distance errors or at a city-level resolution at best " (Laylavi, Rajabifard, & Kalantari, 2016). By applying a geographic context upon a non-geotagged dataset, this study illustrates how the geographic context in text data analysis could help a disaster relief team in understanding how or where people are traveling throughout a disaster thus improving relief efforts.

Some visual analytics tools such as Tableau include GIS mapping tools along with non-spatial data visualization tools. The spatial data analysis functions of GIS can now be used to spatialize and standardize an expansive selection of text data analysis and visualizations. Tableau is a data visualization tool that includes but is not limited to spatial visualization mapping capabilities. Its documentation, "offers unlimited data exploration through an intuitive

interface, encouraging curiosity, creativity, and data-driven decision-making". (Tableau, 2020, Para. 5) Tableau offers the ability to build interactive dashboard; the components in these dashboards can be built out of any of Tableau's multiple types of visual analytics. For the sake of this study, we will be using Tableau Desktop so that we can implement dashboard components

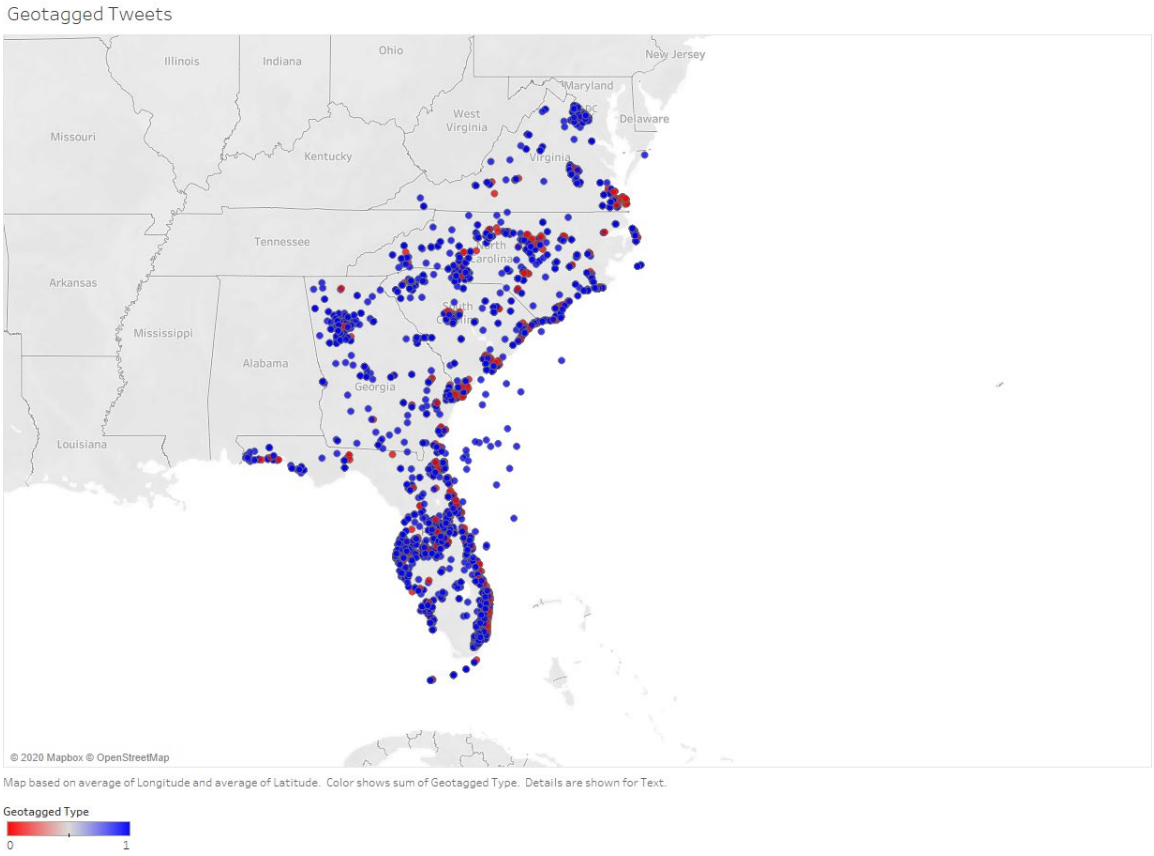


Figure 3. Twitter Data Spatial Visualization Example in Tableau

capable of both spatial and non-spatial visualizations. Figure 3 is an example of spatial visualization using Tableau with the geotagged Twitter data from the dataset. In the graph, the tweets are divided by which attribute from the Twitter API was used to record the location of the device publishing a given tweet. The blue color indicates Twitter users using a preexisting location in Twitter, while the red color indicates a Twitter user using their exact geographic coordinates during publishing of a given tweet. For example, this visualization shows that an overwhelming majority of geotagged Tweets reference preexisting geographic places within the

Twitter application.

Dashboard

The dashboard is essential to provide a clear and concise platform to share results for any users of the application. In order to build a Dashboard to allow the user to navigate through the different results of analysis, the python framework Dash will be compatible with the rest of the python application, hosted within an Django web application environment. Dash can be used alongside Python web framework Django, and data visualization framework Plotly.js in order to build data science web applications (Dash 2020, Para. 2). While ArcGIS will be the primary spatial and temporal visualization tool in our software application design, Tableau will be the software used for the analysis and dashboard designs in this study. Tableau is advantageous for dashboard development because it has the ability to juxtapose varying methods of visual analysis. While ArcGIS is exclusive to GIS, Tableau is a more diverse suite of visual analytics tools that will better fit the entirety of text data analysis conducted in this study.

The overall goal of the Dashboard is to organize components of the different conducted text analysis in a way that aids the user in understanding the disaster. These components will provide the visualizations previously described as well as written information about each analysis. Tableau is advantageous software for building dashboards because it allows for both static and dynamic layouts. With a dynamic layout, Tableau dashboards have the ability to "show/hide content like web pages or images, not just Tableau worksheets…" which is advantageous because "Keeping content on one screen can maintain context better than switching tabs." (Tableau, 2020, Para. 8) This idea of designing a dashboard to represent a user interface on a website allows for the user to compare different analysis by having it all in one

place. In a static dashboard, the user would be have to switch from sheet to sheet without the ability to compare different visualizations in the same viewing.

Proposed Framework

In summary, the proposed framework will consist of a text data analytics dashboard built out of Tableau desktop. All of the data cleaning and pre-processing will be conducted in an Anaconda environment using Python and the Python data science framework Pandas. By using Pandas and the given environment, the software program will be able to take an input of JSON files containing Twitter data and outputs a whichever text data analysis techniques are selected by the user. These scripts will separate the text data analysis of the Tweets from a separate analysis of the Twitter user object in order to provide complementary results about trends in the Tweets themselves and the users posting them within the dataset. Using Pandas will also allow for an output of CSV file which can than be imported and visualized within Tableau. A production environment would require CSV files in uniform of column headings and data types. Since this project is focused on the disaster context and a complimentary dashboard design, this study will use the technique of importing separate Tableau workbooks in order to consolidate the

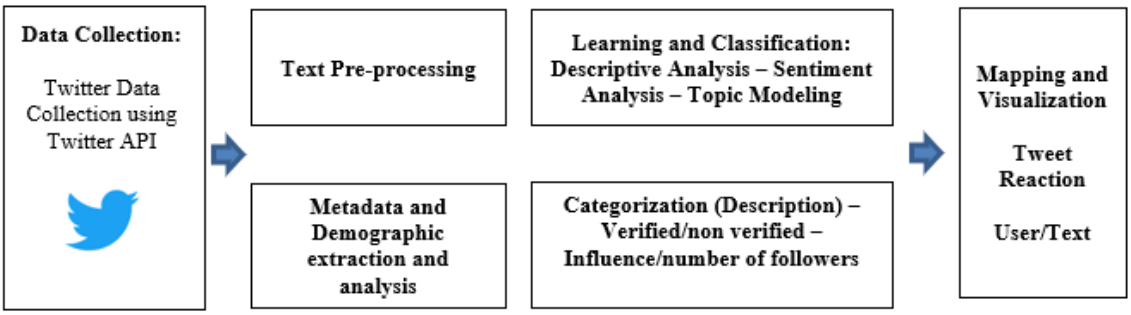


Figure 4. Proposed Framework for Text Data Processing, Analysis, and Visualization

variety of data required for the given dashboard components. Figure 4, above, is an activity flow of the proposed framework.

CHAPTER 4: RESULTS

Weather Advisory Dataset

In addition to the dataset from Twitter, it was important to collect the data regarding Hurricane Dorian's physical path and relative wind speeds. This information was collected from the National Oceanic and Atmospheric Administration database, available on the National Hurricane Center's website (National Oceanic and Atmospheric Administration). This meteorological data is important in contextualizing Hurricane Dorian within the disaster

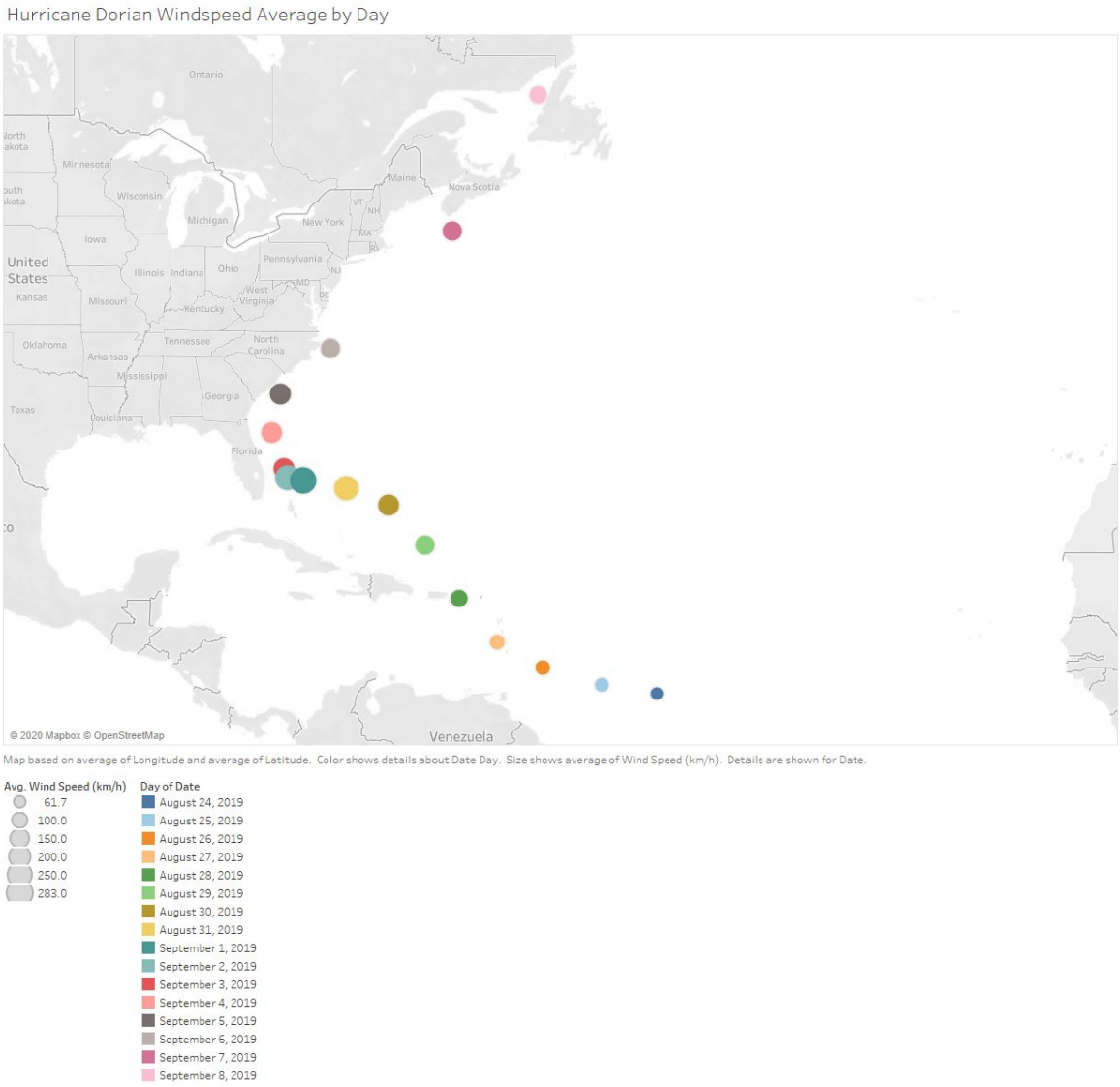


Figure 5. Hurricane DORIAN Advisory Archive

lifecycle. Having a visualization of the highest sustained wind speeds and the physical path of

the storm provides a temporal and spatial context when understanding the Twitter dataset. The visual above could serve purpose on a dashboard as a component that illustrates which regions are experiencing most severe weather conditions. This aspect of the disaster context could help guide users in knowing whether a region is in the preparation, survival, or recovery phase of the disaster lifecycle. This dataset provides temporal and spatial context to Twitter datasets particular to meteorological disasters.

Pre-processing

In order to prepare the JSON dataset for analysis, the pre-processing phase began with importing all of the individual JSON files into Pandas using pandas built-in methods. Once the JSON data was tabulated into DataFrames, we were ready to begin the cleaning process. In order to efficiently clean the data, the two focuses were on cleaning the text of the tweet itself and extracting necessary metadata from the user information.

For cleaning the text, the first filter that was applied onto the dataset was to filter all of the Tweets that were written in the English language using the 'lang' attribute of a Tweet object. While stanfordNLP has the ability to analyze other languages, the scope of the current study is limited to tweets written in English. Not including the tweets in the dataset that were not written in English; would be one limitation and can be addressed in future research projects. Filtering Tweets based on just the 'lang' attribute did not remove all non-English Tweets since there is a group of non-English speakers whose account metadata indicates the English language. It could be argued that the inability to remove all of the tweets not written in English could affect the integrity of the results; however no words from other languages were found in the analysis' results. Another step in filtering the tweets was to exclude retweets or replies. In order to understand the trace of information, we believed it would be most efficient to focus on original tweets (root tweets) so that we could then measure the degree to which the tweet performed in

regards to retweets, replies, and favorites. When the Twitter data was collected; however, the tweets that had measurements of retweets, replies, or favorites, were only the tweets that had been retweeted, replied, or favorited at the time of collection. Therefore, we had to collect a separate collection of the root tweets for these interactions by extracting Tweet objects out of the *retweeted_status* attribute for non-original tweets. This attribute contains a representation of the tweet that was retweeted. By having a separate collection of these retweeted objects, there was adequate data to be able to measure information flow.

At this point, we were ready to start cleaning the text of the individual tweets. While emojis are an element of the Twitter text that can often complicate text data analysis since they are not alphabetical characters (Ayvaz & Shiha, 2017), it was decided that evaluating emojis was beyond the scope of the project since our focus was evaluating geotagged data and the disaster context. Therefore, we removed all emojis from each tweet along with any special characters. The cleaning method then proceeded to lemmatize the words in a tweet to consolidate the number of unique words within the dataset. This is important so that the same word is not evaluated separately due to differences in capitalization.

We employed Tableau for spatial visualization and analysis. While the Tweet objects in JSON contain the information necessary for analysis, it was essential to tabulate this information efficiently within DataFrames in Pandas so that we could export the data to CSV format, which is accepted by Tableau. After cleaning the tweets, additional methods were utilized to extract both metadata on the user account as well as geotagged data for spatial visualization and analysis.

The user metadata was straightforward since all Twitter User objects have the same attributes. The geotagged data; however, is much less uniform due to the Twitter user's

individual choice to designate a physical location at the time of publishing the tweet. When evaluating which geographic attributes could be efficient in visualizations, it was decided that the 'place' and 'coordinates' would be the most efficient due to containing exact geographic coordinates of the device at the time of publishing the tweet. We investigated including the user object's location attribute to increase the dataset's size; however, we found that users too commonly assign their location to an abstract or non-physical location. Additionally, the user location would not represent the victim's physical location within a disaster that would not benefit the study's goal. Because we are hoping further to understand the behavior of victims within a disaster, it was imperative to ensure that the tweet's location was within range of the event. For example, a user's profile location could have been from California; however, if they published their tweet from Florida during the Hurricane we could confirm that they were physically present for the event. This also is important in verifying news agencies from across the country whom would have physically been reporting from the Hurricane's path.

After preprocessing, the total dataset consisted of 698, 931 unique Tweets published by 315, 100 unique Twitter users. Of the dataset, 549,380 tweets contained geotagged data and 11,621 of those tweets were published in either Florida, Georgia, South Carolina, North Carolina, or Virginia. Of the dataset, there were 212, 647 unique tweets that had been retweeted at the time of collection.

In the next few sections, we report the analysis of the total dataset and the geotagged dataset. Specifically, we will perform metadata analysis as well as word frequency, sentiment analysis, and topic modeling analysis.

Twitter Dataset Description

Users

Users were categorized by two separate metadata attributes: verification and followers. In the verification attribute, each individual user account has a Boolean property indicating whether the Twitter account was verified through twitter or not. In these case, verification is a process conducted by Twitter to ensure whether or not a Twitter account is actually the individual or organization that it claims to be. This is most commonly used among accounts for businesses or public figures such as journalists or celebrities. As shown in the figure above, an overwhelming majority of Twitter users are unverified; this dataset contained 291,168 more unverified users than verified users.

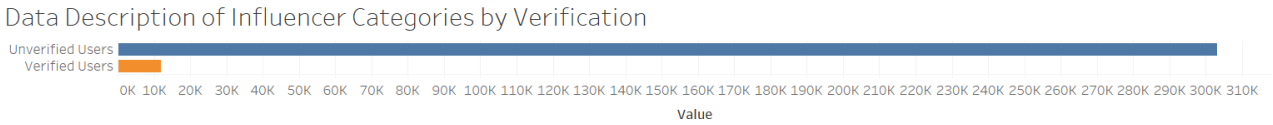


Figure 6. Influencers Description by Verification

The other attribute of user metadata indicating influence is the number of followers of a given account. As shown in the diagram above, an 90% of the users in the dataset have under 5000 followers which is expected for Twitter accounts of personal use.

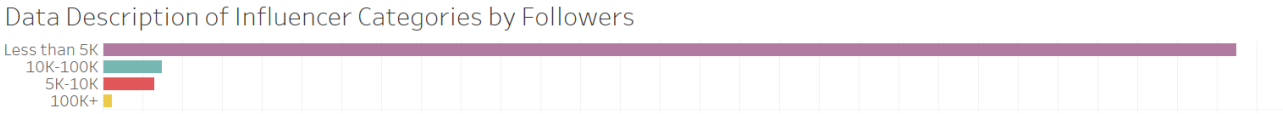


Figure 7. Influencers Description by Amount of Followers

Number of Followers Distribution

90.39%	Less than 5K
4.13%	Followers5K-10K
4.71%	Followers10K-100K
0.77%	Followers100K+

Table 2. Distribution of Influencers by Followers

One significant thing about these distributions is that while accounts with over 100K followers have more weight in an individuals' feed; it actually only makes up less than 1% of

users in our particular dataset. This bar chart illustrates that the overwhelmingly majority of users on twitter will have higher exposure to tweets from users who fall within smaller groups of having higher numbers of followers. While this is not indicating of particular tweets' influence in this study, it does highlight one statistical perspective of tweets influence based on which account publishes a given tweet.

Most Popular Users

Most Popular Users

Screen Name	Name	Description	User Location	User Follow..	Verified	Created At
'realDonaldTrump'	'Donald J. T..	'45th President of the United States of Americaus'	'Washington, DC'	63,742,410	True	3/18/2009 1:46:38 PM
'cnnbrk'	'CNN Breaki..	'Breaking news from CNN Digital. Now 55M strong. Check @cnn fo..	'Everywhere'	55,739,046	True	1/2/2007 1:48:14 AM
'nytimes'	'The New Y..	'News tips? Share them here: http://nyti.ms/2FVHq9v\n\n"The We..	'New York City'	43,998,945	True	3/2/2007 8:41:42 PM
'CNN'	'CNN'	'It's our job to #GoThere & tell the most difficult stories. Join us! F..	None	42,685,059	True	2/9/2007 12:35:02 AM
'BBCBreaking'	'BBC Breaki..	'Breaking news alerts and updates from the BBC. For news, featur..	'London, UK'	40,480,456	True	4/22/2007 2:42:37 PM
'NASA'	'NASA'	'Explore the universe and discover our home planet with @NASA. ..	None	32,512,326	True	12/19/2007 8:20:32 PM
'Pink'	'P!nk'	"it's all happening"	'los angeles'	32,146,770	True	4/4/2009 1:16:34 AM
'BBCWorld'	'BBC News ..	"News, features and analysis from the World's newsroom. Breaki..	'London, UK'	25,694,328	True	2/1/2007 7:44:29 AM
'HillaryClinton'	'Hillary Clin..	'2016 Democratic Nominee, SecState, Senator, hair icon. Mom, Wi..	'New York, NY'	25,296,079	True	4/9/2013 6:04:35 PM
'TheEconomist'	'The Econo..	'News and analysis with a global perspective. Subscribe here: http..	'London'	23,996,629	True	5/12/2007 1:04:50 PM

Name broken down by Screen Name, Name, Description, User Location, sum of User Followers, Verified and Created At.

Figure 8. List of Most Popular Users

When analyzing the most influential user accounts, we looked at the user metadata of the ten Twitter accounts with the most followers. Shown above, the collection of Twitter users consists of the following breakdown:

Journalism/Media	6
Politician	2
Government Agency	1
Celebrity	1

Table 3. Categorization of Top Ten Most Popular Users

It was not surprising to find that all of the accounts were verified as well as assigned a user location to a metropolitan area, excluding NASA who did not have a specific location assigned to the account. It is also understandable that all of the accounts would be verified since these are organizations and individuals that are well known globally.

Word Frequency

We analyzed word frequencies in the total dataset and the geotagged dataset. Comparing word frequencies between the total dataset and geotagged dataset exhibited that the geotagged data excluded words related to controversial statements made by Donald Trump. By looking at the word frequencies by day, we can see that this app was most commonly mentioned after the storm had left the region on September 8, illustrating people's reactions to the preparation phase of the disaster lifecycle. Figure 10 illustrates a word cloud of the top 15 word frequencies of the

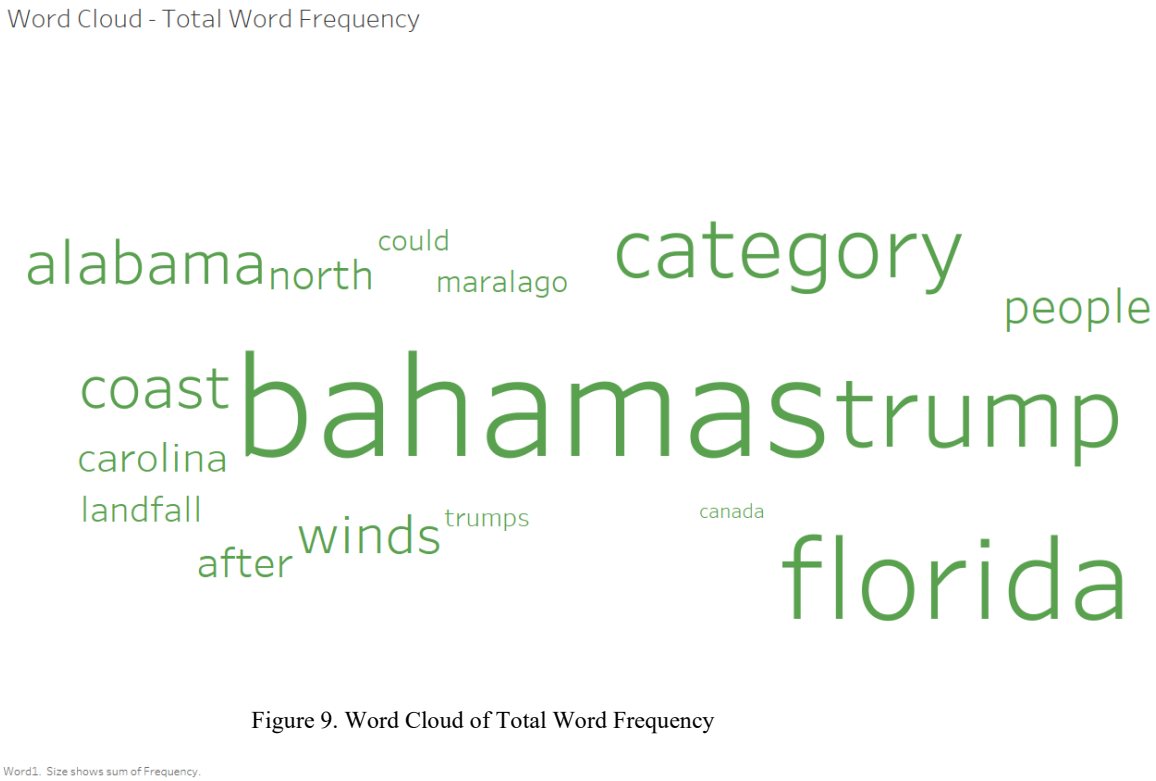


Figure 9. Word Cloud of Total Word Frequency

entire dataset while Figure 11 illustrates a word cloud of the top 15 word frequencies of the geotagged dataset.

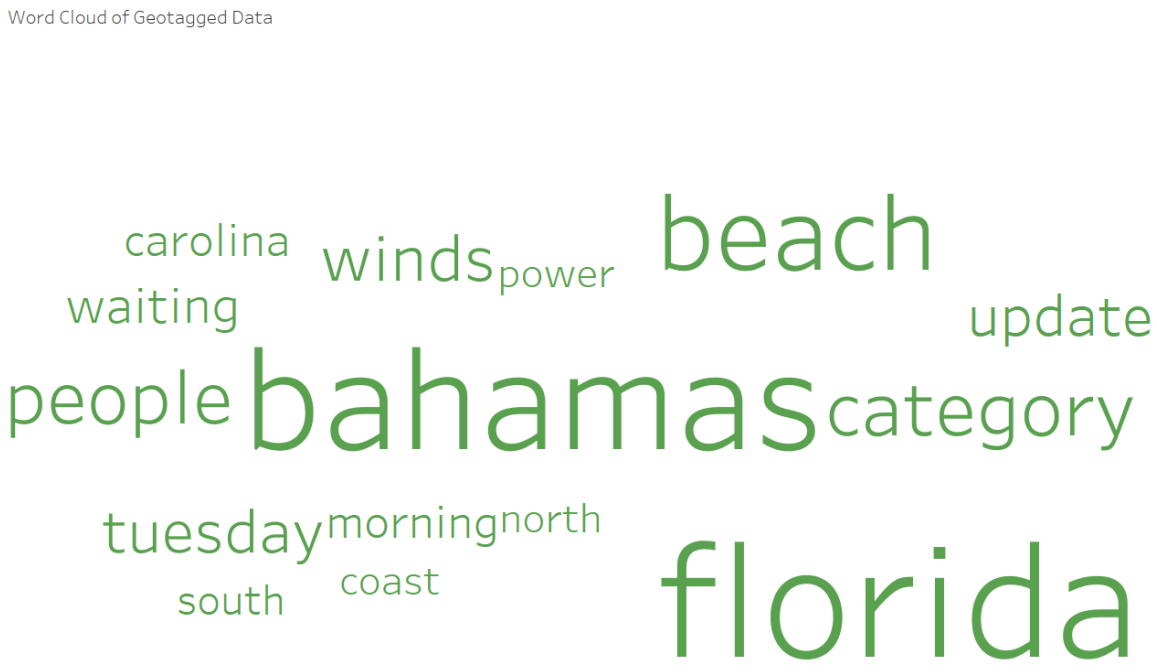


Figure 10. Word Cloud of Geotagged Word Frequency

When using the word clouds to compare the total dataset and the geotagged dataset, we found that there was no presence of the word 'trump' in the geotagged dataset. We also find an absence of maralago in the geotagged dataset which indicates that the word frequencies from the geotagged dataset were less related to the way people feel about Donald Trump or his leadership throughout the event. This finding reiterates the idea that the geotagged dataset maintained less political commentary than the total dataset. Words like 'update' and 'Tuesday' are also stronger related to the nature of the hurricane and its victims' experiences which is more beneficial for the disaster context.

Providing a temporal context to word frequency results helps visualize a relationship between specific events of the disaster and the most frequent words in the dataset. This is one benefit to the timestamp attribute of the Tweet objects since LDA does not provide any temporal

context. This would pair well with LDA output on a dashboard component because it would help analysts identify reactionary trends to the information flow provided by media outlets and government leadership. Figures 12 and 13 show linear breakdowns of word frequency over the days of the event of the total dataset and geotagged dataset, respectively.

For example, the timeline in Figure 12 shows the discussion of Trump on Twitter directly

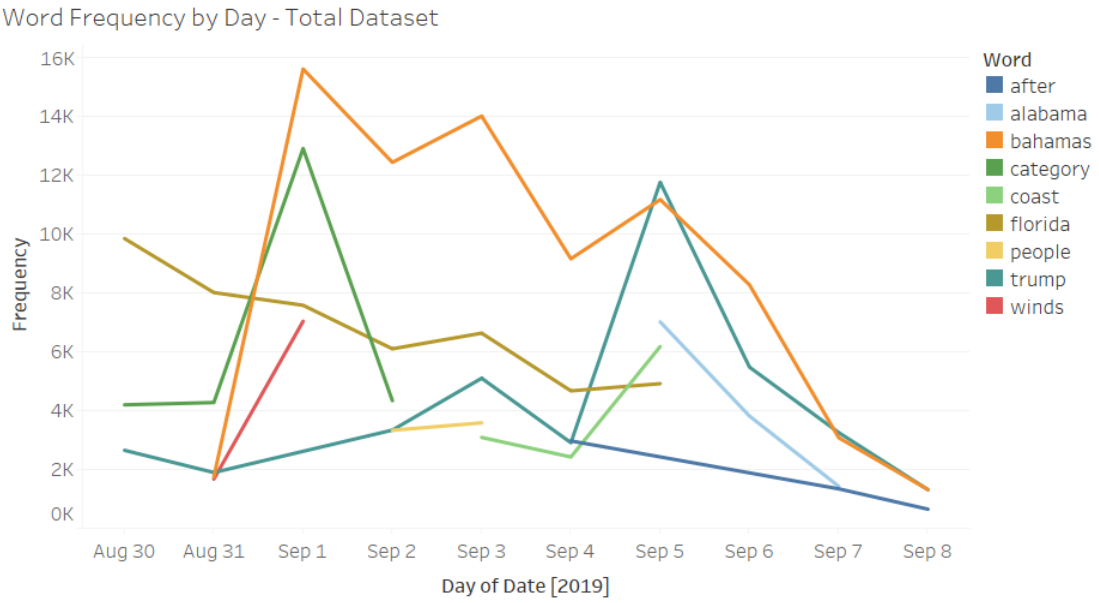


Figure 11. Total Dataset Word Frequency by Day

in relation to the days in which his statements are made and reported on. This is indicated by the spike in 'trump' frequency on September 5 a day after 'Sharpiegate' followed by the presence of 'trumps' September 7 -8 after his comment on the Bahamian death toll the morning of September 7. The geotagged data; however, It also helps illustrate the disaster lifecycle as terms like *Bahamas and category* spike on September 1 as the Bahamas transitioned from survival mode to recovery mode while *Florida* has an increase in frequency between Sep. 2 – 3 which is when the state enters survival mode due to Dorian's physical proximity. We also see *after* present from Sep. 7 – 8 which indicates users going through a transition from the survival phase to the clean-up phase of the disaster lifecycle.

The temporal context of word frequency could benefit a dashboard component by providing a real-time update of the most frequently used words used surrounding the event. Visualizing these frequencies by day could help victims and relief agents in putting together a timeline of the event and seeing which phase of the disaster lifecycle a victim is experiencing. Filtering is also important in that these temporal contexts word filtered only for words which occurred on one more than one day. By focusing on the word frequencies that occur on more than one day, we have a stronger ability to visualize the trends in word frequency over time.

Latent Dirichlet-Allocation

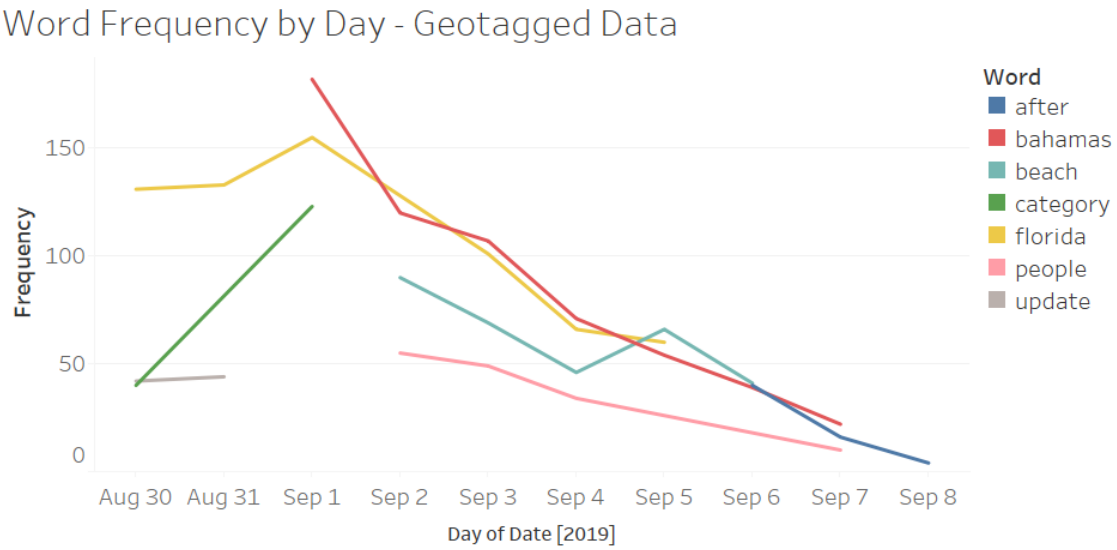


Figure 12. Geotagged Dataset Word Frequency by Day

We find a distinct pattern when looking within LDA's output of topics generated from the entire Twitter dataset. Using the total dataset, we were able to find the following topics conducting LDA from 6 topics, and 6 sub-topic words: (1) Outer Banks damage/relief efforts, (2) geographic updates of hurricane path, (3) Donald Trump claim on death toll in Bahamas, (4) disaster update for victims, (5) North Carolina damage and power outage, and (6) Hurricane Dorian-Alabama controversy 'Sharpiegate'. The full topics and their sub-topic words can be found in the Tables section of the Appendix.

One difference between the LDA output and the geotagged dataset was the frequency of 'colortherapyapp' found in the word frequencies for the geotagged dataset. When researching 'colortherapyapp', it was discovered that this application is "commonly used to relax individuals with different shades of colors." (ColorTherapy, 2020).

Two-thirds of the topics generated are related to outcomes of the event experienced by disaster victims, while one-third of the topics is related to a specific incident in information flow. These results highlight two distinct incidents of misinformation, or inaccurate information, from United States President Donald J. Trump throughout Twitter while the controlled number of topics was equal to six. Topics 1, 2, 4, and 5 were all considered advantageous because they all were specific to a geographic location by name and all were related to victim awareness in the hurricane. Topics 3 and 6 were considered less advantageous to the disaster context because they were related to the political nature of the government's involvement in disaster relief. Therefore, Topics 1, 2, 4, and 5 were represented as a stronger representation of what the victim is experiencing and using to make informed decisions.

Topic 3 represents a tweet from Donald J. Trump's official Twitter account that said "without the help of the United States and me, their would have been many more casualties" (The Independent, Para. 2, 2019) when discussing the death toll in the Bahamas entering recovery. This statement brought controversy due to the fact that he made the announcement before the death count had been completed and was viewed as premature by Bahamian leadership who believed "hundreds and perhaps thousands [were] still missing." (The Independent, Para. 3, 2019) This topic specifically highlights a controversial moment of information flow in which leaders of different countries disagree on the outcome of disaster survival and are perpetuating opposing views of the disaster into the same platform.

Trump's efforts of information flow are present by the present of subtopic word 'alabama' in Topic 6. The incident known by the public as 'Sharpiegate' began when "Trump tweeted about Dorian threatening Alabama on 1 September, apparently relying on information that was several days old." (The Guardian, para. 12, 2019) on September 4, he then released a video from the oval office where he featured a photo of the path of the hurricane that was also a few days old at the time. The billboard had an older prediction of Dorian's path with a circle drawn by sharpie entering alabama. Afterwards, the National Weather Service from Alabama came out in opposition of Donald Trump's statements saying that Alabama would not be of tremendous danger of the path of Hurricane Dorian, this time highlighting disagreement between the president and a weather agency in his own government. Ultimately an official from the NOAA came out in support of the president and in opposition to Birmingham's weather center. This incident was noted to bring attention to the potential politicization of information flow throughout a disaster. One journalist from NPR wrote, "Underlining the reaction by meteorologists to the escalating debate over the president's claims is the fear that weather forecasting itself is becoming politicized." (NPR, Para. 10, 2019)

Understanding this type of information flow is important in understanding where disaster victims are gaining information and illustrates more of the conversation present within social media in a disaster context. Topic 2 illustrates subtopic words referencing the geographic path of the disaster and the weather agency NOAA, while Topic 3 has multiple references to the president and his significant claim. This juxtaposition of sources could benefit a dashboard component because it illustrates the different topics of information that are the most present within the text corpus. This juxtaposition could be beneficial in understanding trends in

politicization of disaster events and how misinformation is spreading throughout disaster victims' conversations.

Using the geotagged dataset, we were able to find the following topics conducting LDA from 6 topics, and 6 sub-topic words: (1) geographic path of hurricane, (2) preparation phase of disaster lifecycle, (3) sentiment of disaster participants, (4) disaster warnings and updates, (5) anticipation of US landfall of storm, and (6) disaster path updates.

The geotagged dataset differs from the entire dataset, by having less evidence of Donald trump gaslighting and had a larger percentage of topics regarding the physical event of the disaster. This is important in the design of a dashboard for disaster relief because it shows that there could be an advantage in geotagged data to filtering out text data that is political in nature or less relevant to the events of the disaster lifecycle. It also illustrates that comparing geotagged data to non-geotagged data could help illustrate external events of the disaster that could potentially misinform the disaster's participants. One disadvantage to the geotagged dataset is the repetition of subtopic words like 'bahamas' and florida' in topics 1, 2, 3, 5, an 6. This could be hypothesized due to the fact that there is a large percentage of geotagged data from Florida. This could be improved; however, by increasing the size and variety of the geotagged data if possible. A larger dataset could also increase the chance of having more detailed subtopic words as found in the LDA output of the entire dataset. The biggest significance between using LDA between a total dataset and a dataset of exclusively geotagged data is the detection of topics related to gas lighting efforts from president Donald Trump.

Sentiment Analysis

A sentiment analysis component is important to a dashboard in visualizing the feelings disaster victims are experiencing based on a region's phase of the disaster lifecycle. After pre-processing, each tweet was evaluated for sentiment using stanfordNLP's Sentiment Treebank in model in which the tweet is given a sentiment name and sentiment value from 0-4. The

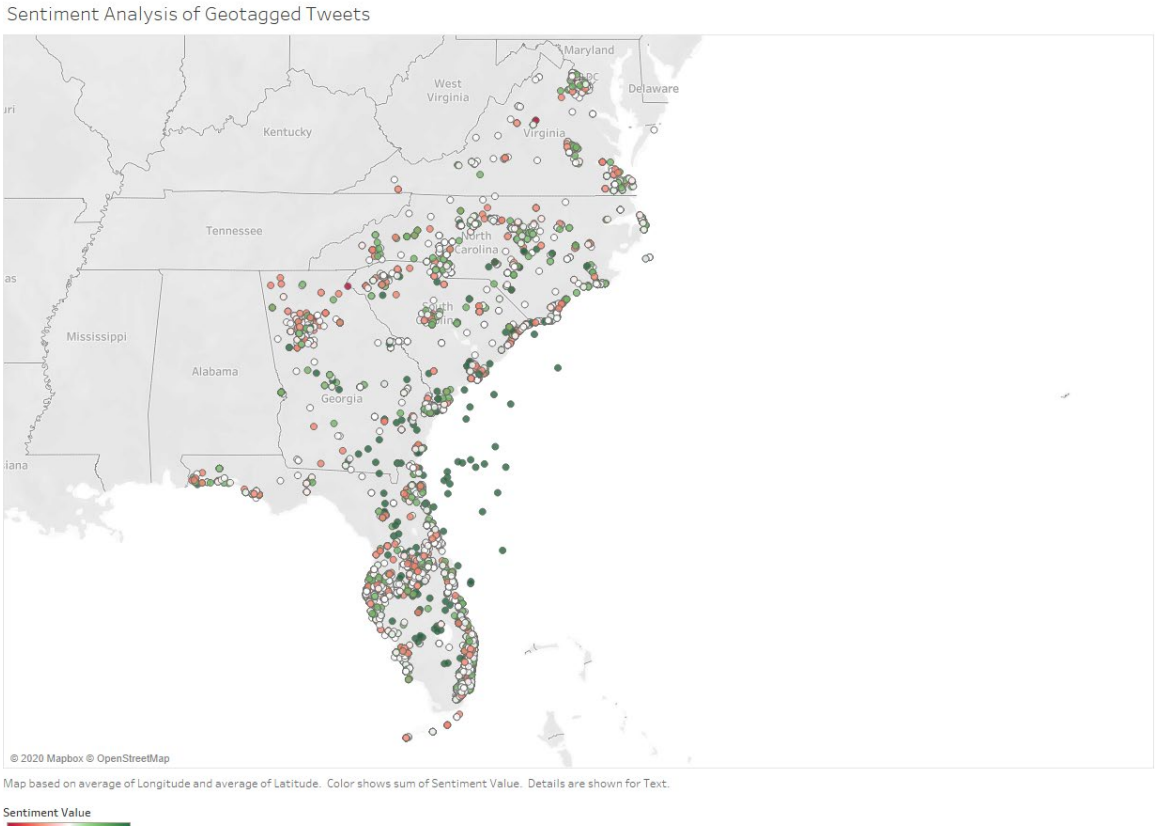


Figure 13. Sentiment Analysis of Geotagged Tweets

sentiment value score starts with 0 being very negative and ends with 4 being very positive.

In the figure above, we can see larger clusters of negative sentiments found in Florida than other states. This could be deduced as a result of Florida receiving the strongest category of Hurricane Dorian throughout the United States. Another trend found in visualizing the sentiment analysis results are larger cluster of positive tweets indicated by green that are found throughout the non-coastal regions of the states. This pattern could also be suggested as a result of being further away from the hurricane and having less damage than the coastal regions which have

higher clusters of neutral and negative sentiments. A temporal context to sentiment analysis would help breakdown which regions are experiencing which phase of the disaster lifecycle. The table below

has examples of tweets from the dataset, and their sentiment and sentiment value.

Tweet	**Sentiment**	**Sentiment Value**
this is awful	Very negative	0
i cannot even imagine being stranded in this	Negative	1
aiken county is at elevated risk of high winds according to the national weather service current forecasts call	Neutral	2
dear father in heaven please be with these people they are suffering much from dorians rage	Positive	3
absolutely stunning	Very Positive	4

Table 4. Sample Tweets from Sentiment Analysis

When looking at the total breakdown of the sentiment analysis, we found the exact same distribution of sentiments between the total dataset and the geotagged dataset. We decided that comparing the two datasets would not benefit the study since the overall sentiment breakdown did not change. The ability for spatial and visualizations; however, still proved valuable for the geotagged dataset as this context of the sentiment analysis could be beneficial in understanding the sentiment of regions in disaster relief efforts.

We also looked at individual visualizations of the sentiment analysis per day of the hurricane. Regions saw higher clusters of negative and neutral sentiments on the days in which a given region was closer to the physical path of the hurricane. As a region would enter the clean-

up phase of the lifecycle, we found an increase in clusters of positive sentiments. Overall the temporal analysis of sentiment would be strongly improved with a larger collection of geotagged data.

Dashboard

When building the dashboard, we imported the separate Tableau workbooks into one larger workbook. By doing so, we were able to build dashboard components out of the previous analysis while maintaining separate data sources. This was advantageous because this dashboard mockup was for design purposes; however, in the case of a production level dashboard, there would need to be one cohesive data source in Tableau since one can only build a dashboard from the same data source. Multiple data sources can be connected with a union in Tableau; however, they must be uniform in their column headings. In order to allow for different visualizations to be accessed in the same container, we created functional buttons that will hide or show different components when clicked. The following images show the different available components and their interactions with individual buttons. The weather advisory dataset was chosen to be static

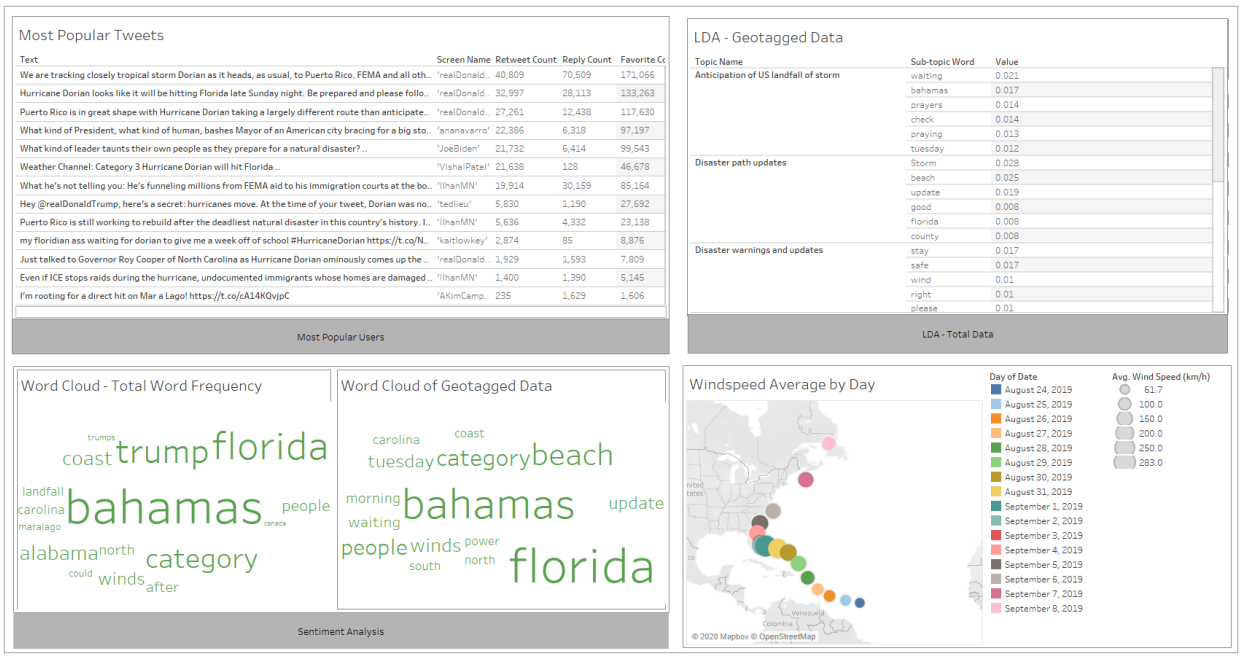


Figure 14. Tableau Dashboard A

since a user could benefit from comparing the path and strength of the hurricane when reading other visualizations.

After choosing a four-quadrant design for the layout of the components, we than decided which components would be most optimal in a disaster context. The lower right quadrant is the only component present in both views since the location and time of the hurricane event is an important context in understanding the surrounding components' analysis. By using the buttons in the dashboard, the user can toggle between seeing different comparisons of the total dataset to the geotagged dataset. For example in the lower left quadrant, the two word clouds are side by side while the upper right quadrant has the option to toggle between LDA results of the two datasets. Figure 16 shows the other components available; however, any combination of available components can be viewed with usage of Tableau's button components.

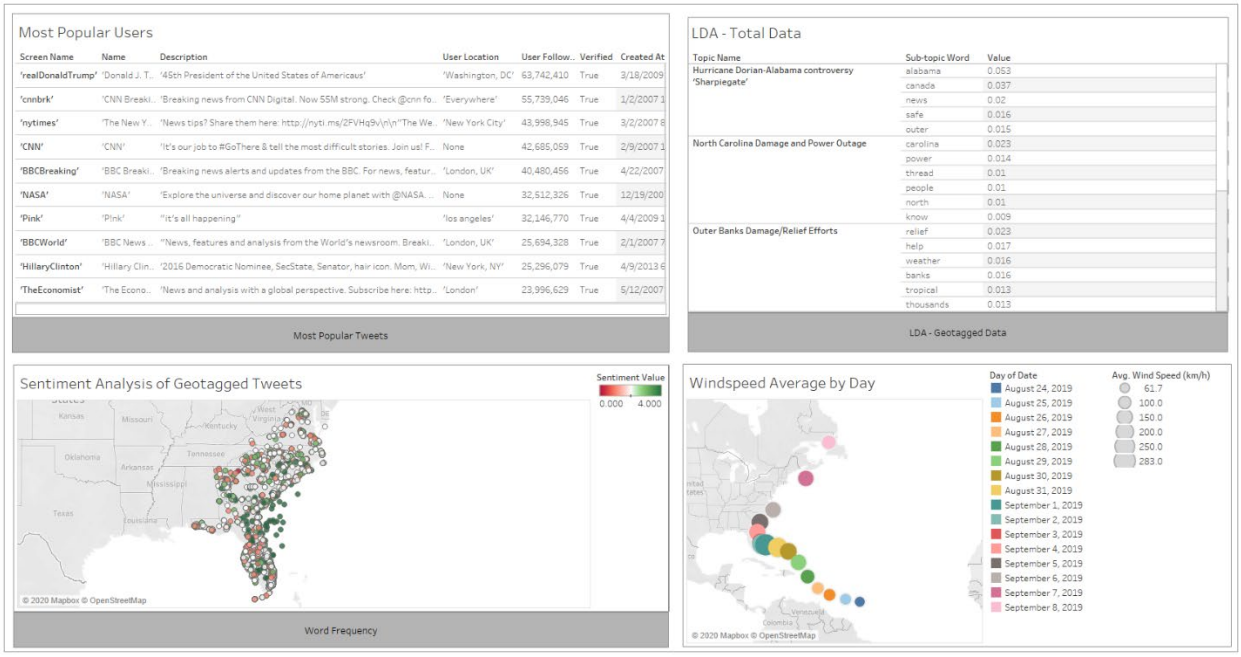


Figure 15. Tableau Dashboard B

Most Popular Tweets

In order to further understand the flow of information on Twitter throughout the hurricane, we looked at the tweets that consisted of the ten most retweets, replies, and favorites respectively. Interestingly, this came down to fourteen unique tweets that consisted of the highest number of reactions for each category. Four of the tweets were published by Donald Trump which is consistent with the metadata from our most popular users, since Donald Trump had the highest number of followers of any user within the dataset. Unlike the gaslighting found from our LDA technique, the most reacted tweets of Donald Trump's were updates on the storm consistent with disaster organizations like FEMA. It is important to recognize that ten of the tweets were published in between August 26 – 30 which indicates that the highest incidents of

Most Popular Tweets

Text	Screen Name	Retweet Count	Reply Count	Favorite Count	Day of Created At
We are tracking closely tropical storm Dorian as it heads, as usual, to Puerto Rico. FEMA and all oth..	'realDonaldTrump'	40,809	70,509	171,066	August 28, 2019
Hurricane Dorian looks like it will be hitting Florida late Sunday night. Be prepared and please follo..	'realDonaldTrump'	32,997	28,113	133,263	August 29, 2019
Puerto Rico is in great shape with Hurricane Dorian taking a largely different route than anticipate..	'realDonaldTrump'	27,261	12,438	117,630	August 29, 2019
What kind of President, what kind of human, bashes Mayor of an American city bracing for a big sto..	'ananavarro'	22,386	6,318	97,197	August 28, 2019
What kind of leader taunts their own people as they prepare for a natural disaster? ..	'JoeBiden'	21,732	6,414	99,543	August 28, 2019
Weather Channel: Category 3 Hurricane Dorian will hit Florida ..	'VishalPatel'	21,638	128	46,678	August 28, 2019
What he's not telling you: He's funneling millions from FEMA aid to his immigration courts at the bo..	'IlhanMN'	19,914	30,159	85,164	August 28, 2019
Hey @realDonaldTrump, here's a secret: hurricanes move. At the time of your tweet, Dorian was no..	'tedlieu'	5,830	1,190	27,692	September 7, 2019
Puerto Rico is still working to rebuild after the deadliest natural disaster in this country's history. I..	'IlhanMN'	5,636	4,332	23,138	August 26, 2019
my floridian ass waiting for dorian to give me a week off of school #HurricaneDorian https://t.co/N..	'kaitlowkey'	2,874	85	8,876	August 27, 2019
Just talked to Governor Roy Cooper of North Carolina as Hurricane Dorian ominously comes up the ..	'realDonaldTrump'	1,929	1,593	7,809	September 5, 2019
Even if ICE stops raids during the hurricane, undocumented immigrants whose homes are damaged ..	'IlhanMN'	1,400	1,390	5,145	August 30, 2019
I'm rooting for a direct hit on Mar a Lago! https://t.co/cA14KQvjpC	'AKimCampbell'	235	1,629	1,606	August 28, 2019

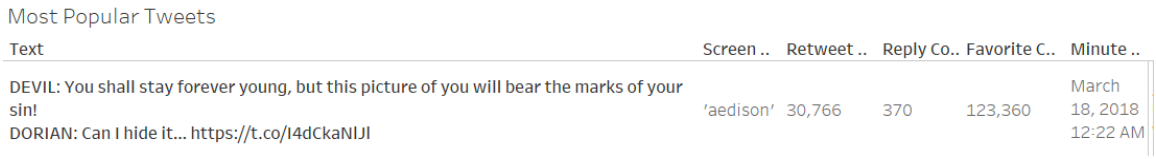
Figure 16. Table of Most Popular Tweets

amplification occurred during the preparation phase of the disaster lifecycle. This also illustrates a higher rate of Twitter engagement in the days before the storm. This could be partly due to the fact that there were large incidents of power damage as well as other destruction during the survival phase which could have had a negative effect on Twitter usage during those days.

It was also observed that 70% of the most reacted tweets came from politicians, only one of which was not a U.S. politician. The caveat to this trend would be that both Donald Trump and Ilhan Omar had multiple tweets within this group. It is also noted that the tweets from Ilhan

Omar, the U.S. Representative from Minnesota, discuss political events surrounding Donald Trump's leadership throughout the disaster. In fact, all but three tweets of the group that were not issued by Trump directly criticized Donald Trump's leadership throughout the event.

An outlier of this group in the initial results was the third most retweeted and favorited tweet which actually was published before Hurricane Dorian, and presumably was a joke about the novel The Picture of Dorian Grey. This tweet references the negotiation between the novel's protagonist Dorian and the Devil, where Dorian receives immortality in exchange for the ability to see his own reflection. This finding shows how the name of the Hurricane itself can


Most Popular Tweets

Text	Screen ..	Retweet ..	Reply Co..	Favorite C..	Minute ..
DEVIL: You shall stay forever young, but this picture of you will bear the marks of your sin! DORIAN: Can I hide it... https://t.co/I4dCkaNIJI	'aedison'	30,766	370	123,360	March 18, 2018 12:22 AM

Figure 17. Most Popular Tweet - Irregular Tweet Sample

compromise the amount of information on Twitter that is helpful to the disaster context. This is most likely to occur if the name is also found in another entity of popular culture. We discovered that this tweet would have been recirculated throughout the #HurricaneDorian hashtag in order to be collected within this dataset and this finding highlighted the importance of using temporal filtering when looking for the tweets with the highest engagement. Initially, this collection was filtered based purely on engagement; however, this finding illustrated that filtering on the 'created_at' attribute is also necessary when applying this type of analysis to a disaster context.

Density Map

One trend that was identified using density maps of twitter activity, showed that the geotagged tweets from verified Twitter accounts were closer in proximity to the physical path of the hurricane than the overall collection of tweets. It could be suggested that is partially in relation to verified accounts from media outlets that broadcast closer to the storm's location. This trend could benefit a dashboard design by possibly helping identify a physical path while also prioritizing verified accounts in information flow since the geotagged results have been more accurate in depicting the events of the disaster. The physical path of verified users is an area that could be further explored in understanding if verified accounts produce more accurate tweets than unverified accounts since their physical path more resembles the path of the hurricane than the dataset as a whole.

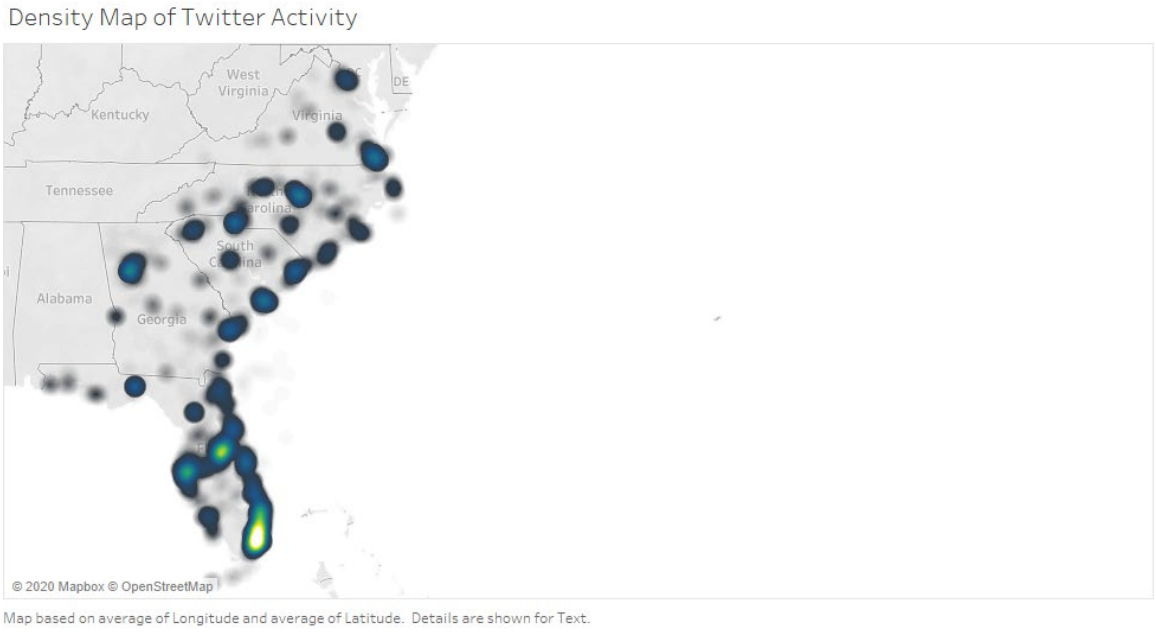


Figure 18. Density Map of Twitter Activity

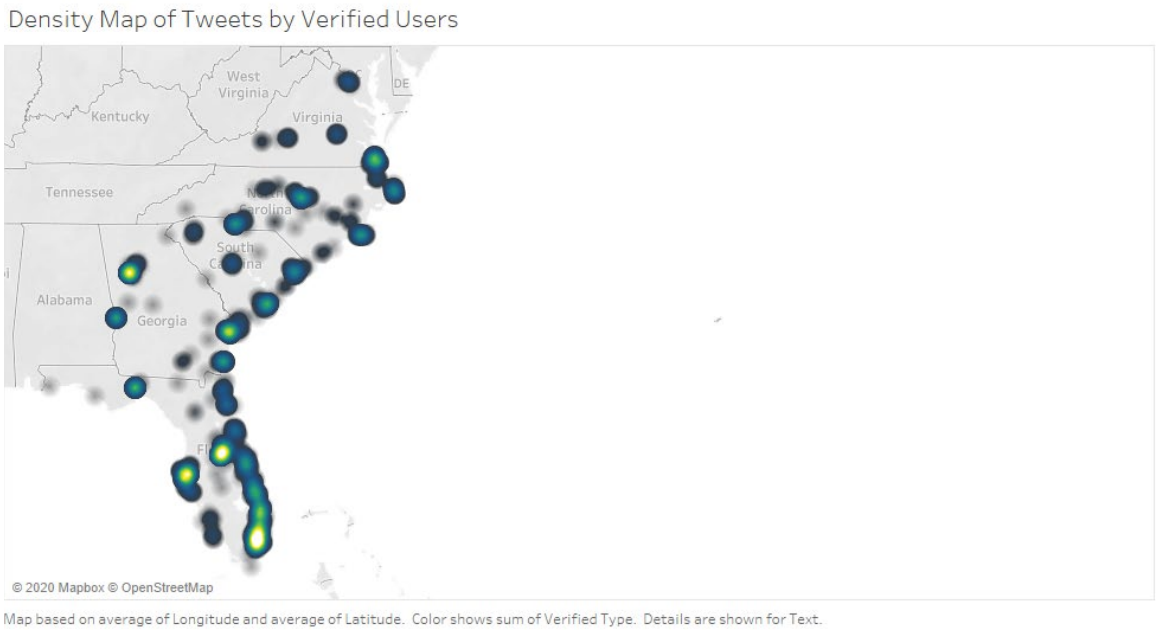


Figure 19. Density Map of Twitter Activity from Verified Users

Lastly, we compared the previous density maps to a density map of all of the tweets within the dataset that were retweeted at the time of collection. When using the retweet count attribute as our measurement variable, we found a very distinct outline of all of the major metropolitan areas in the states that experienced the hurricane. The hottest part of the map paralleled our previous density maps since the majority of retweets came from southern Florida, the region which took the most damage. Understanding that the majority of retweets came from major cities helps visualize the information flow and the amplification occurring to tweets in big cities despite the fact that not all of the major cities were as close to the physical path of the storm. This result could also be due to the fact that Twitter accounts from news organizations would more likely be based out of these cities. This density map of retweets provides a visual analysis for where retweets are published which further explains the information flow throughout the disaster.

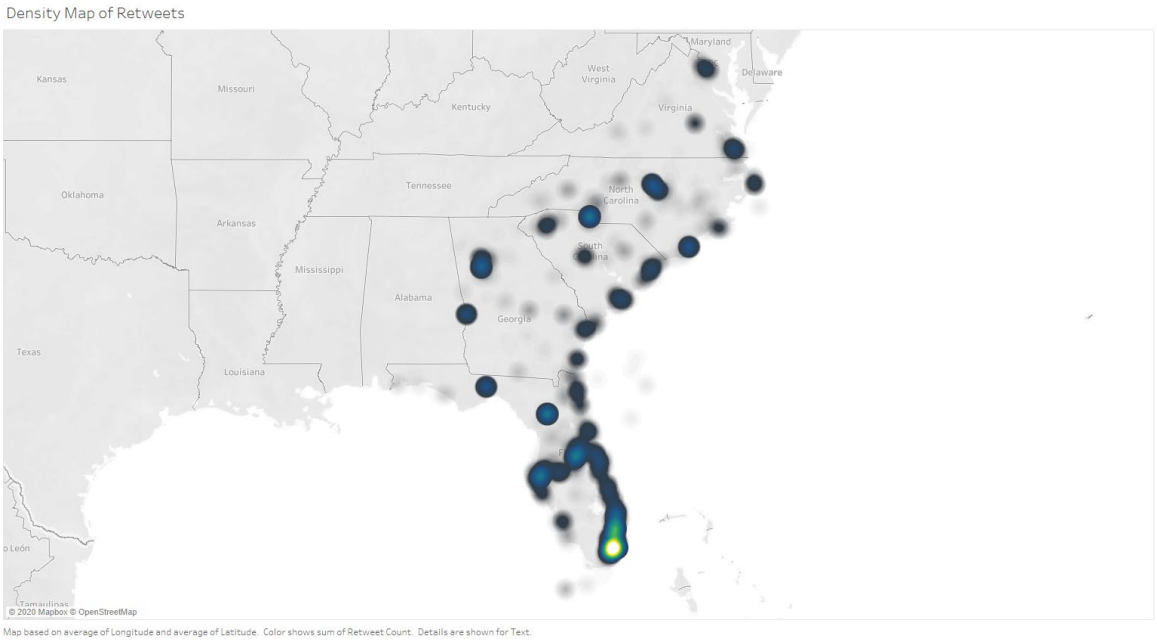


Figure 20. Density Map of Retweet Activity

Using the density map context for geotagged data proved to be beneficial in understanding the flow of information throughout Twitter during the disaster. Most specifically, the density maps of verified activity and retweets were effective in outlining different ways in which news and media outlets may report on a disaster event. While the verified density map has the ability to show verified profiles of individuals within media, the retweets activity than shows exactly where the verified tweets and others are actually being amplified to other disaster victims. Using the two density maps is advantageous to looking at the overall geotagged dataset since it outlines a specific type of user and behavior, respectively.

CHAPTER 5: DISCUSSION

Text data is abundant and ever growing as the usage of social media platforms like Twitter are continually growing and becoming more integrated into our society. The disaster context is one that is new for these platforms and will continue to grow in available data, as more disasters occur. The relationship between a Twitter user and their physical location can be one way of filtering through the noise that occurs during an event. In our case of Hurricane Dorian, this relationship was advantageous in the pursuit of understanding where disaster victims learn about the event around them. The nature of a hurricane is advantageous for this type of study because it has a physical path over time. The hurricane also occurs in the fashion that can be measured since there are distinct phases of the disaster lifecycle such as preparation, survival, and clean up. Other disasters like earthquakes that are less predictable, may prove less advantageous in this type of temporal analysis. Other types of disasters like a pandemic, whose physical path is more abstract, may need additional measures in order to efficiently understand its relationship to behavior on Twitter.

As Twitter also becomes a larger platform for politicians, the flow of information throughout a disaster becomes more apparent throughout these collections of tweets. However, as this study highlighted, this does not indicate that all of the most amplified messages from politicians throughout a disaster are related to efforts of disaster relief. This type of filtering will be beneficial to helping future disaster victims and should not be overlooked by disaster relief agencies. Dashboard designs are also efficient tools that should be utilized in disaster relief efforts. Software applications like Tableau that do enable automatic updates based on newly obtained data can also be efficient in juxtaposing different visualizations. Enabling the user of a dashboard to customize the analysis and visualizations they read will ultimately give disaster relief members more freedom in finding the best means to help victims. Text data from

platforms like Twitter will never be able to fully cover all victims' experience due to the fact that economic disadvantages will prevent certain victims from being able to engage on these social media platforms. It should not be considered an end all be all to understanding victims experiences, but is a complimentary tool for populations that are engaged in these platforms.

The most surprising finding in these results has been the overwhelming abundance of political figures, political discussions, and political contextualization of the hurricane throughout the dataset. When we were in the beginning phases of processing, we underestimated the amount of politicization that occurs to any event in the U.S. Donald Trump has been well known for his strong presence on Twitter; however, it could be argued whether the amplification of his messages throughout Hurricane Dorian was due is due to his persona before holding office or his current role as president of the United States. Regardless, it was somewhat surprising that he would have a distinct lead in influence over other news organizations or media outlets throughout the platform. This pattern was also significant in finding the most popular tweets, where the overwhelming majority of the tweets were from politicians about political circumstances of the disaster.

Understanding the amount of political discourse that occurs on Twitter is essential to developing adequate filters to get through the unrelated noise. Without these detailed filters, the noise and hysteria that comes with political discourse bleeds into the data analysis. Too much political influence could render the components in a dashboard useless which is why adequate filtering is essential to the success of the dashboard designed in this experiment.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

When comparing geotagged data to a dataset consisting of both geotagged and non-geotagged data, this study concluded that the geotagged data contained higher frequencies of text related to the physical events of Hurricane Dorian. The geotagged dataset produced text analysis that could be argued as being more effective in understanding disaster victims' experiences. Expanding this study of isolated geotagged data would be one aspect of the project which could be continued to further evaluate whether geotagged data is more accurate in covering a disaster than non-geotagged data. Implementing statistical models to detect trends in misinformation could benefit disaster victims in making decisions throughout a disaster. It also could benefit disaster relief agencies in tracing information flow among victims while potentially providing an explanation for unusual trends or reactions throughout groups of people. This type of tool within a dashboard could become more useful for future disasters if there continues a trend of increased politicization among natural disasters or mainstream media.

Another aspect of this research project which could be continued in further work would be to expand the software application design to include collecting, cleaning, and visualizing the weather advisory dataset whether through webscraping or access to an API from the National Hurricane Center. By integrating this data which was collected manually and processed using Pandas, the software application could enhance visuals and provide a wider variety of analysis with the additional information. It should also be noted that the production requirements for Tableau to have uniform CSV files would be one aspect of the project that could be enhanced further. Because the metadata analysis and text data analysis require such a wide range of attributes, it could be viewed as an obstacle to design the whole database of the system around this uniformity. This could also be an issue when trying to run a production environment since

there would not be the convenience of importing different workbooks as conducted in this experiment.

The ultimate goal of this research is to be able to identify trends among the Hurricane Dorian dataset between characteristics of the users and a given sentiment or topic, as well as between users' location and the location of the hurricane. This journal also has provided a series of text analysis visualizations that could benefit a dashboard component which could provide disaster relief agencies with a visual opportunity to find trends between the text data and the physical path of a given natural disaster.. It is important to see how we can improve the ways we understand the mental processing of victims of a natural disaster in order to be able to improve the way we help given victims. By hopefully identifying patterns of information processing specific to this hurricane, this journal hopes to produce a framework design that can be used for analyzing Twitter data for future natural disasters. While this journal specifically focuses on the path of the eye of Hurricane Dorian, the end goal is to be able to visualize the path of any natural disaster's geographic coordinates next to visualizing the analysis of geotagged Twitter data. Hopefully by pairing these two side by side, analysts will be able to deduce further trends in a given area, resulting in a more optimized rescue effort.

There is also a chance that by emphasizing a disaster relief context, we may identify alternate relationships between types of users and quantitative trends that may be applicable to other contexts of opinion mining. By hopefully understanding a trend in the way disaster victims value ethos among Twitter data, relief agencies will hopefully be able to improve the way that messages are exchanged between victims and relief team members to hopefully be more effective in their rescue efforts. The ultimate idea with this software is that relief agencies can make better informed decisions when they can see who is providing the most trustworthy

information and what the trends are in an area at a given distance from the source of the natural disaster. This study's contributions hope to inspire further research into text data analysis among disasters of all kinds, not solely meteorological. By contributing a dashboard design for disaster monitoring, an overview of text data analysis methodologies, and an evaluation of current technologies in geographic information systems, we hope this study will further enhance ways in which social media data can be applied to disaster relief efforts.

REFERENCES

1. Ayvaz, S., & Shiha, M. O. (2017). The Effects of Emoji in Sentiment Analysis. *International Journal of Computer and Electrical Engineering*. https://doi.org/10.17706/ijcee.2017.9.1.360-369
2. Beigi G., Hu X., Maciejewski R., Liu H. (2016) An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief. In: Pedrycz W., Chen SM. (eds) Sentiment Analysis and Ontology Engineering. Studies in Computational Intelligence, vol 639. Springer, Cham
3. Bi, R., Schleier, M., Rohn, J., Ehret, D., & Xiang, W. (2014). Landslide susceptibility analysis based on ArcGIS and Artificial Neural Network for a large catchment in Three Gorges region, China. *Environmental Earth Sciences*. https://doi.org/10.1007/s12665-014-3100-5
4. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research. https://doi.org/10.1016/b978-0-12-411519-4.00006-9
5. Bhayani, Go, and L. Huang. "Twitter Sentiment Analysis". 2009.
6. Bravo-Marquez, Mendoza, Poblete. Meta-level sentiment models for big social data analysis. 2014.
7. CNN. Hurricane Sandy Fast Facts. https://www.cnn.com/2013/07/13/world/americas/hurricane-sandy-fast-facts/index.html
8. Color Therapy. The perfect relaxation App through social coloring. https://www.colortherapy.app/. Accessed 2 October 2020.
9. "Introduction to Dash." *Plotly | Dash*, 2020, dash.plot.ly/introduction.
10. "Documentation." *ArcGIS for Developers*, developers.arcgis.com/documentation/. Accessed 28 November 2019
11. Hu, X., Tang, L., Tang, J., & Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. https://doi.org/10.1145/2433396.2433465

12. JSON. http://www.json.org/. Accessed 18 November 2019.
13. Kaji, N., & Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of HTML documents. EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.
14. Laylavi, F., Rajabifard, A., & Kalantari, M. (2016). A multi-element approach to location inference of Twitter: A case for emergency response. *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi5050056
15. National Oceanic and Atmospheric Administration. National Hurricane Center and Central Pacific Hurricane Center. Hurricane DORIAN Advisory Archive. https://www.nhc.noaa.gov/archive/2019/DORIAN.shtml?
16. NPR - National Public Radio. NOAA Contradicts Weather Service, Backs Trump on Hurricane Threat In Alabama. Published 6 September 2019.
17. Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly: Management Information Systems*. https://doi.org/10.25300/MISQ/2013/37.2.05
18. Pang and L. Lee. "Opinion Mining and Sentiment Analysis" in Foundations and Trends in Information Retrieval, 2008.
19. Pasco Sherrif on Twitter. https://twitter.com/pascosheriff/status/906712903868469249?lang=en. Published 9 September 2017.
20. Kryvasheyeu, Y., Chen, H., Moro, E., Van Hentenryck, P., & Cebrian, M. (2015). Performance of social network sensors during Hurricane Sandy. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0117288
21. Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. Journal of Statistical Software. https://doi.org/10.18637/jss.v091.i02
22. Schulz, A., Thanh, T. D., Paulheim, H., & Schweizer, I. (2013). A fine-grained sentiment analysis approach for detecting crisis related microposts. *ISCRAM 2013*

Conference Proceedings - 10th International Conference on Information Systems for Crisis Response and Management.

23. Snopes. Does Hurricane Florence Now Contain Sharks? https://www.snopes.com/fact-check/hurricane-now-contains-sharks/. Published 7 September 2017.
24. Tableau. Create Dynamic Tableau Dashboard Layouts with Sliding Containers. https://www.tableau.com/about/blog/2016/1/how-create-dynamic-tableau-dashboard-layouts-sliding-containers-48269. Accessed 29 October 2020.
25. Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P. (2011). User-level sentiment analysis incorporating social networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2020408.2020614
26. The Guardian. Top weather official who defended 'Sharpiegate' makes tearful clarification. https://www.theguardian.com/us-news/2019/sep/10/noaa-trump-alabama-sharpiegate-statement-clarify. 10 September 2019.
27. The Independent. Trump says Hurricane Dorian death toll in Bahamas would be higher without his help, as bodies still being counted. https://www.independent.co.uk/news/world/americas/us-politics/trump-tweet-bahamas-death-toll-hurricane-dorian-latest-a9095686.html. Published 7 September 2019.
28. Twitter. Docs - Twitter Developers. https://developer.twitter.com/en/docs Accessed 18 November 2019.
29. Utz, S., Schultz, F., & Glocka, S. (2013). Crisis communication online: How medium, crisis type and emotions affected public reactions in the Fukushima Daiichi nuclear disaster. *Public Relations Review*. https://doi.org/10.1016/j.pubrev.2012.09.010
30. Wang, Z., et al. "application of Gis Rapid Mapping Technology in Disaster Monitoring." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-3, 2018, pp. 1849-1851.

31. Qi, W., & John E., T. (2014). Quantifying, Comparing Human Mobility Perturbation during Hurricane Sandy, Typhoon Wipha, Typhoon Haiyan. *Procedia Economics and Finance*. https://doi.org/10.1016/s2212-5671(14)00910-1

TABLES

Topic 1 – Outer Banks Damage/Relief Efforts

0.023	relief
0.017	help
0.016	weather
0.016	banks
0.013	thousands
0.013	tropical

Table 5. LDA Total Dataset - Topic 1

Topic 2 – Geographic Updates of Hurricane Path

0.097	bahamas
0.043	storm
0.027	coast
0.025	landfall
0.023	noaa
0.020	north

Table 6. LDA Total Dataset - Topic 2

Topic 3 – Donald Trump Claim on Death Toll in Bahamas

0.089	trump
0.034	trumps
0.024	death
0.022	toll
0.020	president
0.018	claim

Table 7. LDA Total Dataset - Topic 3

Topic 4 – Disaster Update for Victims

0.019	makes
0.017	still
0.016	people
0.014	damage
0.014	good
0.012	left

Table 8. LDA Total Dataset - Topic 4

Topic 5 – North Carolina Damage and Power Outage

0.023	carolina
0.014	power
0.010	north
0.010	thread
0.010	people
0.009	know

Table 9. LDA Total Dataset - Topic 5

Topic 6 – Hurricane Dorian-Alabama controversy ‘Sharpiegate’

0.053	alabama
0.037	canada
0.020	news
0.016	safe
0.015	outer
0.013	made

Table 10. LDA Total Dataset - Topic 6

Topic 1 – Geographic path of hurricane

0.021	florida
0.015	north
0.014	people
0.012	storm
0.012	bahamas
0.009	landfall

Table 11. LDA Geotagged Dataset - Topic 1

Topic 2 – Preparation phase of disaster lifecycle

0.017	dont
0.010	yall
0.010	take
0.009	time
0.008	bahamas
0.008	home

Table 12. LDA Geotagged Dataset - Topic 2

Topic 3 – Sentiment of disaster participants

0.014	help
0.012	florida
0.011	shit
0.010	bahamas
0.008	away
0.008	thing

Table 13. LDA Geotagged Dataset - Topic 3

Topic 4 – Disaster warnings and updates

0.017	stay
0.017	safe
0.010	please
0.010	wind
0.010	right
0.008	still

Table 14. LDA Geotagged Dataset - Topic 4

Topic 5 – Anticipation of US landfall of storm

0.021	waiting
0.017	bahamas
0.014	prayers
0.014	check
0.013	praying
0.012	tuesday

Table 15. LDA Geotagged Dataset - Topic 5

Topic 6 – Disaster path updates

0.028	Storm
0.025	beach
0.019	update
0.008	county
0.008	florida
0.008	good

Table 16. LDA Geotagged Dataset - Topic 6