

**Spatial Text Mining: An Enhanced Text Mining Framework for Extracting Disaster
Relevant Social Media Data**

Christopher Scheele

Advisor: Qunying Huang

A thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science

(Geographic Information Science and Cartography)

at the

University of Wisconsin-Madison

2017

Acknowledgements

First, I would like to express my gratitude and appreciation to my advisor, Dr. Qunying Huang, who gave me the opportunity to work with her on my thesis. With her help and guidance, I could navigate the obstacles of this challenging project. Not only did I develop academically from her, but also personally. Dr. Huang's kindness, patience, and optimism motivated me to see the project to completion. I had a pleasure working with and studying from her. Without her insights and persistent encouragement, none of this thesis would have been possible.

I also extend my thanks to my committee members, Dr. A-Xing Zhu and Dr. Sam Batzli. Working with them over the past few years has been a pleasure. Their expertise in the field of geographic information science was invaluable through the feedback and advice I received during this project. I am grateful for the many skills I have learned from them and hope to apply these skills moving forward in my career.

In addition, I would like to thank members of my cohort, specifically Scott Farley and Starr Moss. They provided me with useful feedback throughout my research and were enjoyable to spend time with outside of work.

I would also like to thank members of the RealEarth™ team at the Space Science and Engineering Center. I appreciate all your technical support for this project.

Finally, to my parents who have supported and encouraged me over the last quarter century. Many thanks and love for being the best parents and life models.

Table of Contents

Abstract	iii
1. Introduction	1
1.1 Research Questions.....	3
1.2 Research Approach.....	5
1.4 Thesis Structure.....	6
2. Literature Review.....	7
2.1 Social Media for Disaster Events	7
2.2 Text Mining for Extracting Disaster Relevant Information.....	10
2.3 Remote Sensing for Tracking Disaster Events.....	12
2.4 Social Media and Authoritative Data Fusion.....	12
3. Spatial Text Mining	14
3.1 Framework.....	14
3.2 Data Streams, Processing, and Storage	16
3.3 Spatial Feature Generation	27
3.4 Annotation and Classification.....	29
3.5 Spatiotemporal Distribution of Sample Data.....	33
4. Experimental Design and Analysis.....	35
4.1 Experimental Design.....	35
4.2 Feature Determination	39
4.3 Feature Combination Performance.....	41
5. Conclusion and Future Study	46
5.1 Conclusion.....	46
5.2 Future Direction	47
References	49

Abstract

In the past decade, the rise in social media has led to the development of a vast number of social media services and applications. Disaster management represents one of such applications leveraging massive data generated for event detection, response, and recovery. To find disaster relevant social media data and automatically categorize them into different classes (e.g. damage or donation), current approaches utilize natural language processing (NLP) methods based on keywords, or machine learning algorithms relying on text only. However, these classification approaches have not been perfected due to the variability and uncertainty in language used on social media. Therefore, more clues or signals are necessary to improve purely text-based approaches. A disaster relevant social media post is highly sensitive to the location and time of the post. Thus, additional features related to space and time could be useful for differentiating relevant posts. However, there has been no systematic study to explore the extent of how spatial features can aid text classification. To fill the research gap, this study proposed a spatial text mining framework to incorporate spatial information derived from social media and authoritative meteorological datasets, along with the text information, for classifying disaster relevant social media posts. This approach assesses the textual content using common text mining methods and the spatiotemporal relationship of the post to the disaster event. An assessment of the framework utilized geo-tagged social media posts and meteorological data for the 2012 Hurricane Sandy disaster event. The study designed and demonstrated how diverse types of spatial features, including wind, flooding, and proximity, can be derived from the data and then used to enhance text mining. Additionally, different temporal features are also derived and integrated into text classification. This study used a common classification scheme for classifying disaster

relevant social media posts into different categories. Commonly used machine learning algorithms, including Naive Bayes and Support Vector Machine classifiers, assessed the accuracy within the enhanced text-mining framework. Finally, integrating textual, spatial, and temporal features to generate different classification models identified the features with the greatest influence in the classification. The experimental results indicate that proximity (spatial), disaster status (i.e., spatiotemporal relationship of the hurricane and social media post) features help improve the overall accuracy of the classification. The results from this study address the need for incorporating spatial data when using social media in disaster management applications.

Keywords: Text mining, Social media, Disaster management, Hurricane Sandy

1. Introduction

In the past decade, the rise in social media has changed the way people interact with each other. Society has access to more information than ever before. Popular social media services, such as Facebook, Twitter, Flickr, LinkedIn, YouTube, and many others are generating data at petabyte or even exabyte levels daily (Huang & Xiao, 2015). As social media integrates with more devices and platforms, various applications develop to harness social media data. Disaster management represents one application that leverages social media to support a variety of activities and operations by disaster managers and organizations. Social media, on one hand, can be used as a platform to provide critical information to the public about hazardous events for relief and recovery efforts (Houston et al., 2014). Disseminating information about a disaster event is helpful when the target audience does not use or have access to standard methods of communication (e.g. telephones, radios, and television). For example, the National Weather Service (NWS) now tweets weather watches and warnings. Disaster managers, on the other hand, can also gather social media data to monitor disaster events in real-time. Citizens involved with the disaster are “sensors” providing geo-located information to supplement authoritative data sources (De Longueville, Annoni, Schade, Ostlaender, & Whitmore, 2010). Having a geographical situational awareness through social media enables the identification of areas with infrastructure damage, affected people, and evacuation zones (Huang & Xiao, 2015).

To identify social media data relevant to a disaster event and extract useful data for disaster coordination and response, different approaches have been developed and used in various applications for free or commercial use (Ashktorab, Brown, Nandi, & Culotta, 2014; de Albuquerque, Herfort, Brenning, & Zipf, 2015; Huang, Cervone, Jing, & Chang, 2015;

Zlatanova, Zlatanova, Holweg, & Holweg, 2004). One of the typical approaches to filter through large amounts of social media is text-match based on searching for specific keywords or groupings of words using machine learning algorithms, known as the text-based approach, to determine if the social media data is relevant to a disaster event (Ashktorab et al., 2014; Bakillah, Li, & Liang, 2014; Huang & Xiao, 2015; Landwehr, 2014). This approach builds a classifier that categorizes text based on the existence of keywords. For example, if a class for damage frequently contained the word “tree” in its training datasets, then a testing post with the word “tree” would more than likely be classified in the damage class. Text-based classification is a fast way to organize large datasets, like social media data.

However, the text-based approach is not without fault. The filtering processes or algorithms cannot correctly identify or differentiate all social media data related to a disaster event based on the contents of text alone due to the variability in language used on social media (Bruns & Liang, 2013). For instance, an algorithm might look on Twitter for the “#hurricanekatrina” hashtag. If a user is unaware of the hashtag or even misspells it, the likelihood of the classifying the data as disaster relevant decreases (Bruns & Liang, 2013).

Additionally, social media data are misidentified by text-based methods when data unrelated to the disaster are incorrectly classified as relevant. For example, during a flooding event, the message “the water is very high right now” could have different contexts. While an individual most likely indicates, “the water level is high” while they are on a riverfront within the disaster area, another individual at home far from the impact region could mean, “the water in their bathtub is high.”

To overcome the challenges of using only text to classify the social media data, classifiers must incorporate other characteristics of the disaster event. Disaster relevant social

media is sensitive to space and time (Huang & Xiao, 2015). Thus, the knowledge of social media posts in space and time is crucial for identifying and interpreting whether a message is disaster relevant. Analysis of other data sources related to the disaster event can discover such knowledge. Given the challenges with current text mining strategies for disaster relevant data, it is important to find an alternative approach utilizing additional data sources, specifically spatial data, to improve the classification accuracy.

1.1 Research Questions

This research aims to improve current approaches to relevant social media data extraction during disaster events. To improve the accuracy of current approaches, this study develops a framework that integrates spatial and temporal information into the text mining classification. Two research questions aid in the development of the new framework.

First, what are the key types of spatial data necessary for classifying social media data during a disaster event? The research question can be broken down into two parts. The first is identifying geographic data from which spatial features can be derived for text classification. A variety of geographic data could contribute to the improvement of social media message classification. Some of the geographic data are highly domain-specific. For example, during a flood event information about how much rain has fallen in each area and lower elevation areas is important. This type of information can be derived from radar data products, weather station observations, and a digital elevation model. However, during a forest fire event climate conditions in the area or wind patterns are more relevant data for the disaster.

Other geographic data are less domain specific, the most prominent being proximity to the disaster. Based on Tobler's First Law of Geography, a person posting on social media about a tornado going through their town is more related to the disaster than a person reading

about the same event on social media hundreds of miles away (Tobler, 1970). In addition to space, time is another important dimension useful in classifying disaster events. For management purposes, a disaster can be broken down into four phases: mitigation, preparedness, emergency response, and recovery. These phases can also be used as general references for social media data (Huang & Xiao, 2015).

The second part of the first research question involves the determination and utilization of the spatial and temporal data as features in the classification models. Current text-based classification models cannot directly assimilate raw data. Instead, the classifier reads features, like text, as coded numeric values. Determining how to code each spatial and temporal feature is an area previously not studied, but vital for this research. After determining the spatial and temporal feature codes, the research addresses optimal utilization of the features with the text data for the classification. With the diversity of spatial data in mind, different combinations of spatial and temporal features were assessed to determine the minimum datasets required to achieve the highest classification accuracy for a given disaster event.

Second, can the inclusion of spatial and temporal information improve the accuracy of current text-mining approaches for classifying social media data? With the spatial data narrowed down from the first research question, different classification models compare the new framework to the current text-only approach. Ultimately, the answer to this question addresses the fundamental research problem, insufficient accuracy of text-only classification for disaster events.

1.2 Research Approach

This research aims to build a spatial text mining framework to retrieve disaster relevant social media data by incorporating geographic information from social media data and the disaster event into the current text-based approaches. Social media as a form of communication during a disaster event is a topic well studied (Houston et al., 2014). Given adequate power and network coverage during a disaster event, people gravitate towards social media to communicate their experiences with the world. However, it is challenging to determine the validity of this type of data based on text alone. Thus, text mining classifiers have problems with accuracy. To provide a promising solution to the accuracy problem, the spatial text mining framework aggregates authoritative spatial data sources, specifically meteorological data. The aggregated spatial data then fuses with the social media data to improve classification.

First, the author collected social media and spatial data, in the form of meteorological data, for the disaster. These two data sources form the backbone of the framework and thus reside in a database for efficient retrieval. After deriving the key spatial features, the meteorological data joins to individual social media posts based on the time and location. Next, additional features are derived based on the temporal information associated with each social media post.

Once the data has been prepared, a random sample of posts are classified manually based on a standard disaster management classification scheme (Imran et al., 2013). Finally, the classified data ran through different supervised machine learning algorithms to classify social media posts based on text only and text plus spatial and temporal data. The expected

results from the framework is that the addition of spatial and temporal data will increase the overall accuracy of the classification of disaster relevant social media data.

1.3 Contribution

There are three contributions of this thesis work. First, this thesis assesses the types of spatial and temporal features necessary for the classification of disaster relevant social media data. Both domain and non-domain specific features are included in the assessment. Second, this thesis demonstrates how to use and integrate spatial and temporal data into text classification for identifying disaster relevant information. Finally, the thesis introduces a framework for fusing spatial data with social media for classification of disaster relevant social media. The framework addresses the shortcomings of utilizing only text to identify and extract disaster relevant social media data when considering spatial data is necessary.

1.4 Thesis Structure

The remainder of the thesis is broken into four sections. Section 2 reviews the literature on the fields related to the thesis. Section 3 describes the methodology and is broken down into six subsections. Section 3.1 provides a description of the framework. Section 3.2 defines the case study and study area. In Section 3.3, a detailed description of all data used and the necessary preprocessing steps is given. Section 3.4 defines the spatial feature extraction algorithm for joining the data. The social media annotation and classification methodology is in Section 3.5. Section 3.6 summarizes the spatiotemporal distribution of the data. Section 4 describes the experimental design and includes subsections for the experiment analysis. Finally, Section 5 summarizes the thesis and discusses potential future directions.

2. Literature Review

2.1 Social Media for Disaster Events

Social media data as a means of communication during disaster events presents many advantages over standard communication methods (Houston et al. 2014). During a disaster event, the ideal communication system is one which is not expensive, simple to use, mobile, and reliable (Mills et al., 2009). Social media services fit this communication system in that they are dependable, available on mobile devices, and can handle high volumes of traffic (Jaeger et al., 2007). Social media also has advantages over traditional communication methods in terms of timeliness of information, relevance at the community level, cost, and adaptability (Kiem & Noji, 2011).

The many advantages of using social media data over traditional communication have led to different use cases in disaster management. One widespread use of social media during disaster events is real-time dissemination of information (Xiao, Huang, & Wu, 2015). Unlike the one-way communication of traditional media, social media allows for both sending and receiving of messages. In the immediate wake of the 2010 Haiti earthquake, studies show that information about the disaster was released through social media services (Kiem & Noji, 2011). The quick dissemination of information on social media allows disaster managers to assess the situation from first-hand accounts and offers a way to communicate with the individuals in the area impacted by the disaster.

Backchannel communications is another way social media informs decisions during disaster events. When traditional sources of communication lack information or cannot keep up to date on current information, backchannel communications represent user-driven information acquisition and sharing parallel to the official data sources (Xiao, Huang, & Wu,

2015). For example, during a California wildfire event in 2007, local news focused their efforts on the impact of the fire on urban areas, but neglected the rural areas. The residents of the rural areas began using social media to share information about the disaster (Sutton et al., 2008). Peer-to-peer backchannel communications on social media fills information gaps when official sources of information are unavailable.

The first two uses of social media during disaster events focus on communication during or after the event takes place. Social media can also detect events related to the disaster. For example, during an earthquake in Virginia in 2011, people in the eastern United States reported learning about the event on Twitter before feeling the earthquake at their location (Ford, 2011). Another form of event detection is finding users in need of help or assistance on social media. For instance, during a 2011 tsunami off the coast of Japan, several tweets were direct requests for assistance (Acar & Muraki, 2011). One of the tweets read, “We’re on the 7th floor of Inawashiro Hospital, but because of the risen sea level, we’re stuck. Help us!” (Acar & Muraki, 2011). This type of message is critical to detect, but requires a high degree of verification (Lindsay, 2011).

To overcome the difficulty of disaster event detection, one of the emerging uses of social media for disaster events is extraction of situational awareness for coordination and relief operations (Huang & Xiao, 2015). For example, Ashktorab et al. (2014), created Tweedr, a Twitter based data mining tool that extracts actionable information for disaster relief workers during natural disasters based on keywords. Tweedr is open-source allowing disaster management at different governmental levels to use the tool.

The use of social media for disasters clearly has a variety of advantages over traditional methods of communication. However, the veracity or uncertainty of the social

media data is one challenge to address. During a disaster event, disaster managers must make timely decisions based on the data available. If the data is unreliable, the decisions could have catastrophic consequences.

Using social media to make decisions during disaster events has raised concerns to the reliability and quality of the data (Goodchild & Glennon, 2010; Goodchild & Li, 2012). The first issue is in the accuracy of the information. Using geo-tagged tweets to find incident locations can pose a problem if the user is tweeting about something they experienced at a different time and location (Gao et al., 2011). Another case of data inaccuracy occurred in 2011 during the Tohoku earthquake. Tweets seeking assistance appeared long after the people in need were rescued creating greater confusion for disaster managers (Lindsay, 2011).

Another problem with the veracity of social media data is when social media is used maliciously (Huang & Xiao, 2015). The generation of social media for pranks, attacks, and rumors is common (Lindsay 2011). Falsified requests for help can draw first responders away from helping those in true need of assistance. Moreover, the rumors and falsified reports can spread through social media with ease (Lindsay 2011).

To tackle the reliability issues associated with social media data during a disaster event, a temporal understanding of the generation of social media from the beginning to the end of the disaster is important. Houston et al. (2014) proposed a simple three-phase disaster classification for social media, pre-event, event, and post-event. During the pre-event phase as previously mentioned social media users send and receive information about the disaster event. Another use during the pre-event phase is to signal or detect disasters. After the disaster event has taken place, social media offers a way to reconnect the impacted

community. For instance, Hurricane Katrina in 2005 caused the displacement of a large population from their homes. Social media enabled displaced residents to share stories and information about the neighborhood prior to returning home (Shklovski et al., 2010). The three-phase classification is a simple way to classify social media data during a disaster event. Other efforts classified social media into the typical four-phase categorization (mitigation, preparedness, response, and recovery) or even forty-seven different themes (Huang & Xia, 2015). However, in a real-time disaster event the sheer volume of data from social media poses a challenge for storing and analyzing the data (Huang et al., 2015). During a disaster event, emergency planners use data collected during the event to make real-time decisions. Users generate social media data at changing rates. Any solution for utilizing relevant social media data during disaster events must have the processing capabilities to handle the stream of data efficiently (Huang et al., 2015). The proposed framework will overcome the challenges of data volume by using text mining techniques to automatically search through the social media data for disaster relevant information.

2.2 Text Mining for Extracting Disaster Relevant Information

One way to overcome the challenges associated with the volume social media data is by searching through the text to look for patterns in the words that might signify data related to a disaster. Disaster managers have applied many different techniques for text mining social media data. The main objective of text mining is to develop a classification scheme or model to predict if a particular social media post relates to the disaster event. The first step in creating a model is generating a set of keywords. With Hurricane Sandy, the keywords might be sandy, hurricanesandy, or hurricanenyc (Huang and Xiao, 2015). These keywords act as an initial filter to remove messages irrelevant to the disaster (Huang and Xiao, 2015). One

inevitable consequence of this filtering approach is not capturing all messages relating to the disaster event. Some social media data users might be unaware of the existence of a certain keyword being used, they might use a unique keyword no one else is using, or their data contains only a picture or video and no text at all (Bruns & Liang, 2013). To improve the overall accuracy, an understanding of the data misclassified as irrelevant in the current research methodology is necessary.

The next step is determining the n-grams used to train the model. N-grams are a set of co-occurring words within a set of words. For example, during a flood a user posts the message “I am stuck in a flash flood please help!” Using the unigram or 1-gram approach means each word becomes a single token read by the classifier. Increasing to a bigram or 2-gram would lead to two word tokens, such as “flash flood” or “please help.” While increasing the amount of words per token yields more information, the classification accuracy does not improve significantly (Halteren, Zavrel, & Daelemans, 2001). Hence, when creating a model for disaster events, unigrams are standard practice (Ashktorab et al., 2014; Spinsanti & Ostermann, 2013; Huang & Xiao, 2015).

Finally, a classification algorithm runs using training data. The four common classification algorithms applied to disaster events include K-nearest neighbors, decision trees, naïve Bayes, and logistic regression (Ashktorab et al., 2014; Bruns & Liang, 2013; Huang & Xiao, 2015; Huang & Xu, 2014; Jain, 2015; Spinsanti & Ostermann, 2013; Xiao, Huang, & Wu, 2015). Improving the accuracy of text mining approaches is the main motivation for this research. Instead of following the current research track of focusing solely on the classification schemes and algorithms themselves, this research incorporates spatial

information about the social media data into the classification algorithm as it relates to the disaster event.

2.3 Remote Sensing for Tracking Disaster Events

Types of disasters, such as hurricanes, floods, or wildfires, remote sensing data provides additional geographic information for detection and tracking to disaster officials during a disaster event. For example, algorithms detect tornadoes by finding slight differences in the patterns of radar images (Alberts et al., 2011). The methods used for detection and tracking of disaster events are similar to the text mining approaches mentioned in the previous section (Roy & Kovordányi, 2012). However, unlike text mining where training data are discrete, remote sensing data for disasters involves training data that are continuous leading to more complex pattern recognition and processing (Lakshmanan & Smith, 2009). Moreover, data mining remote sensing data requires separate identification algorithms and attribute extraction methods for each type of disaster event. In other words, the algorithm to detect and track a hurricane will be vastly different from that of a tornado, whereas generalized text mining algorithms apply to many disasters. Consequently, a high degree of domain knowledge of the disaster in the context of remote sensing is required to accurately detect these types of disasters (Lakshmanan & Smith, 2009). The spatial text mining framework builds upon the current remote sensing data mining methods described by Lakshmanan and Smith (2009) by combining the spatiotemporal information about the disaster with social media data to determine disaster relevant social media data.

2.4 Social Media and Authoritative Data Fusion

During a disaster event, disaster managers and planners use many different data sources to assess the situation. Leveraging other data sources, like satellite or other

geographic data, could improve the analysis of social media data during a disaster event. In previous works, the Twitter data estimated trajectories of earthquakes, tracked the locations of tornadoes, and detected wildfire hotspots (Crooks et al., 2013; Jain, 2015; De Longueville et al., 2010). On the other hand, disaster detection is not always possible using social media as was determined during a 2013 flooding event (Fuchs et al., 2013).

One promising new area of research is to fuse social media data with other forms of spatial data for disaster events. Albuquerque et al. (2015) used “authoritative” hydrological data for a flood event with social media messages to confirm the presence of the flood in the disaster region. Disaster managers supporting relief efforts could then use the messages. They also found that the closer the social media data was to the event, the more likely it was to be about the flood event. This simple quantitative assessment shows how additional datasets can improve social media data identification.

Another approach to fusing remote sensing data with social media is by using the social media data as a way to overcome limitations of remote sensing data. Schnebele & Cervone (2013) used social media data in real-time to verify the presence of water in a specific area during a flooding event when remote sensing imagery was unavailable.

Given the infancy of spatial and social media data fusion for disaster management, no prior work incorporates spatial and text data features for extracting disaster relevant social media data. The next section outlines the spatial text mining framework, including the methodology for both social media and spatial data extraction using data mining algorithms.

3. Spatial Text Mining

3.1 Framework

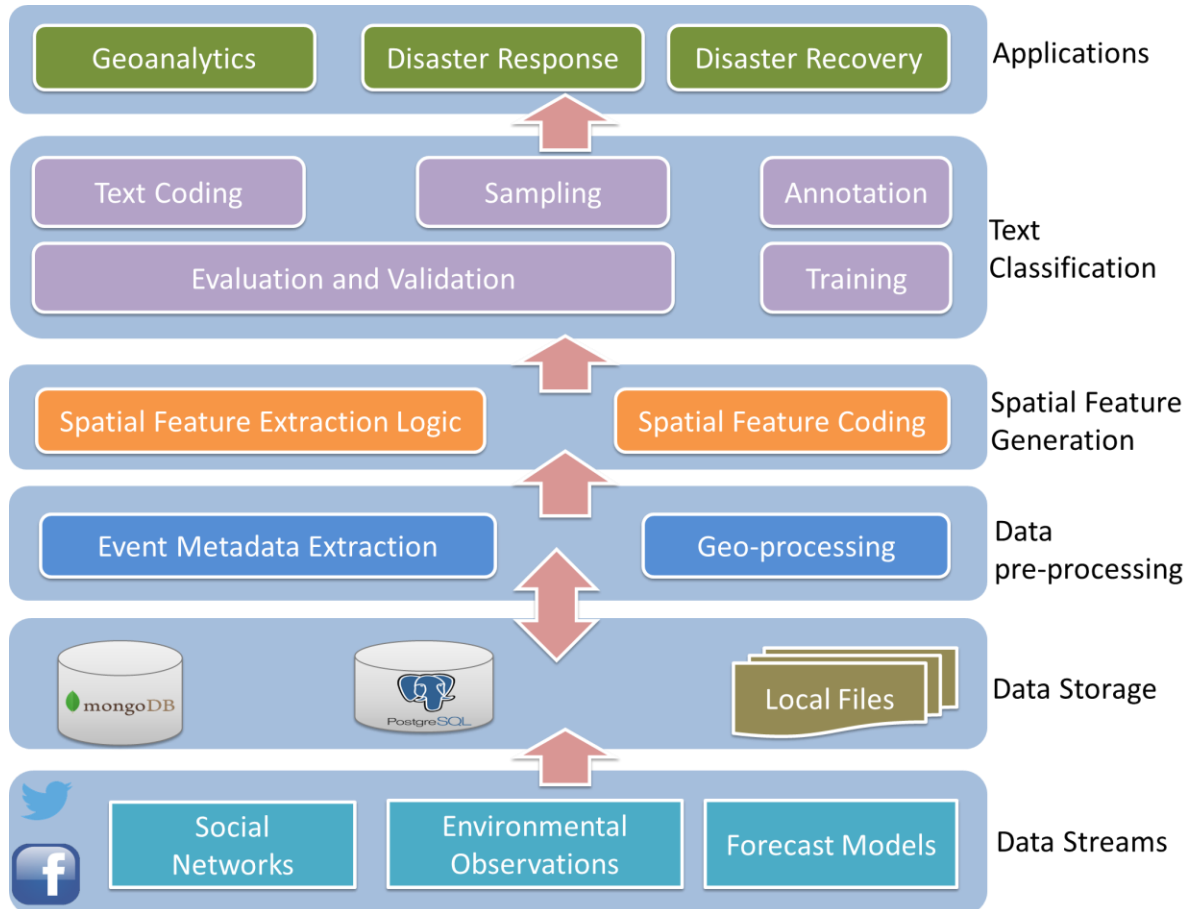


Figure 1. The spatial text mining framework

The spatial text mining framework is an enhanced text mining framework that incorporates situational information about the disaster in the form of geographic data into the supervised text classification. The hypothesis is, by adding information about the disaster during the time of a social media post, the overall accuracy of the supervised text classification will increase. Figure 1 shows a graphical depiction of the framework. There are six key components of the framework:

1. Social media and spatial data streams: A real-time disaster management scenario is the intended use of the framework. Data in a variety of formats (e.g. point, polygon, raster, text) and volumes stream into the framework from multiple sources.
2. Database storage: With the high volume of data flowing into the framework, storage is necessary before and after processing. Due to scalability and spatial functionality, the storage of social media and spatial data is separate. MongoDB can handle the high volume of social media content, while PostgreSQL via the PostGIS extension performs a variety of spatial queries.
3. Event detection and pre-processing: The social media data streams in a uniform format of text with an associated point in time. However, the challenge with the spatial data is its variety. Before the spatial information can be stored, scripts process the data into uniform data types and file formats. This is also the point in the framework where event detection takes place. Disaster event detection is an important step for creating a dynamic spatial filter within the framework compared to a traditional bounding box. In turn, the framework uses spatial information associated with the dynamic assessment of the disaster extent as a feature in the text classification. It is important to note detecting a disaster event using spatial data is an active research topic that is highly domain specific.
4. Spatial feature generation: The center of the spatial text mining framework is the spatial feature extraction. Using fuzzy logic, spatiotemporal information from a social media post generates the spatial features relevant to the disaster. The result is a social media post with metadata in the form of spatial features.
5. Text classification: Staying consistent with current methods, social media data with geographic metadata travel through a supervised classifier to determine disaster relevant

social media posts. This thesis evaluates and validates the results from the classifier to determine if the addition of geographic information improves the text classification.

6. Section 4 introduces the experiment design and analysis.
7. Applications: The front-end interface of the spatial text mining framework is a geanalytics tool for use by disaster managers and planners. The tool allows the viewing of the classified social media data as well as the spatial data. Disaster managers and planners can use the interface to make informed decisions based on insights gained from the data.

3.2 Data Streams, Processing, and Storage

3.2.1 Case study

To test the spatial text mining framework, a disaster case study was selected based on several criteria. First, the disaster had to have taken place within roughly the last decade. Since social media became widely used beginning in the mid-2000s, it is important to select a disaster at a time when a social media platform has developed a large user base. Related to the temporal constraint, the spatial data related to the disaster must be diverse, easily accessible, and quantitative or categorical in nature. For example, an earthquake is a possible disaster to test the framework given the frequency and global coverage. While an earthquake would fit into the framework, the limited amount spatial data associated with disaster does not make it ideal for a first study. Finally, the supervised classification involves training text data for the classifier. In order have the best understanding of the text for training, the disaster must occur in an area where English is the primary language.

A tornado, flood, and hurricane in the United States were the disasters that met these criteria. Other considerations made in selecting the case study include spatial extent,

temporal extent, and impact on the area. Ultimately, Hurricane Sandy from 2012 was selected for the case study due to data availability and previous studies related to text mining for hurricanes.

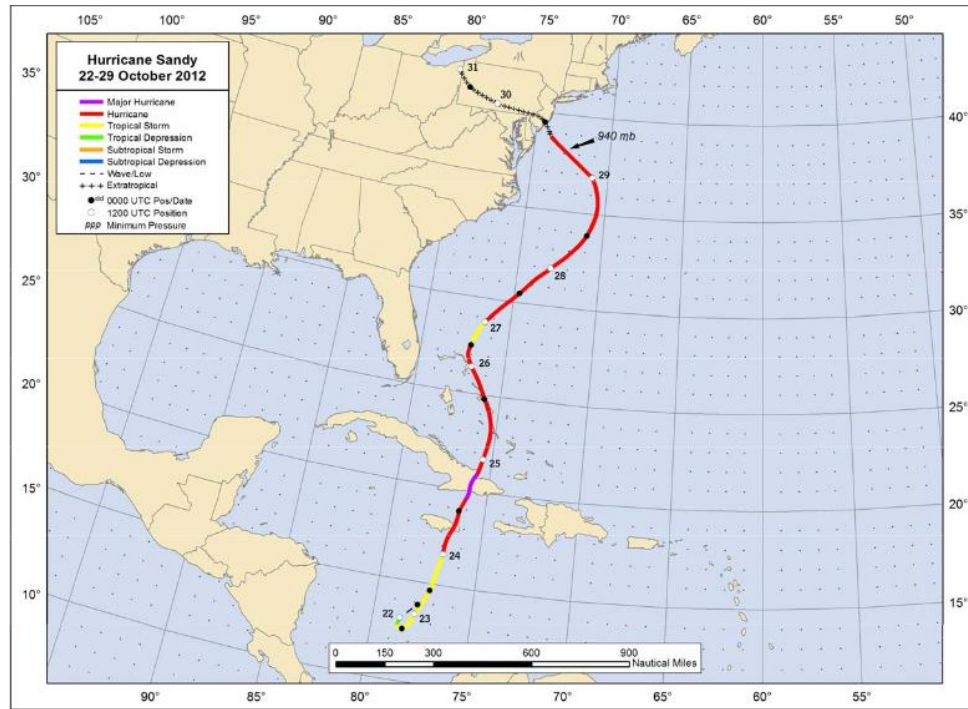


Figure 2. Hurricane Sandy storm track and intensity (Blake E. et al., 2013)

Hurricane Sandy (October 22, 2012 – November 2, 2012) was the 18th named tropical cyclone for the 2012 Atlantic Hurricane. Sandy formed in the south-central Caribbean on October 22 and tracked north intensifying into a hurricane prior to tracking across Jamaica (Figure 2). Sandy quickly intensified to a category 3 storm (wind speeds between 111-129 mph) before striking Cuba, only to weaken back to a category 1 storm shortly thereafter (Blake E. et al., 2013). On October 28, Sandy was a few hundred miles southeast of North Carolina and tracking to the northeast. Concern remained for parts of the Northeast United States as models indicated Sandy would likely turn to the northwest and make landfall in a densely populated area. On October 29 at 23:30 UTC, Sandy made landfall near Brigantine,

New Jersey (just north of Atlantic City) with an estimated intensity of 70 kts winds (Blake E. et al., 2013). Thanks to the cool waters and cold air mass, Sandy made landfall much weaker than predicted. Sandy continued to track northwest into Pennsylvania before the center of the storm became ill-defined midway through October 31 (Figure 2). Over the next two days, the remnants of Sandy merged with a low-pressure area over Canada and disappeared.

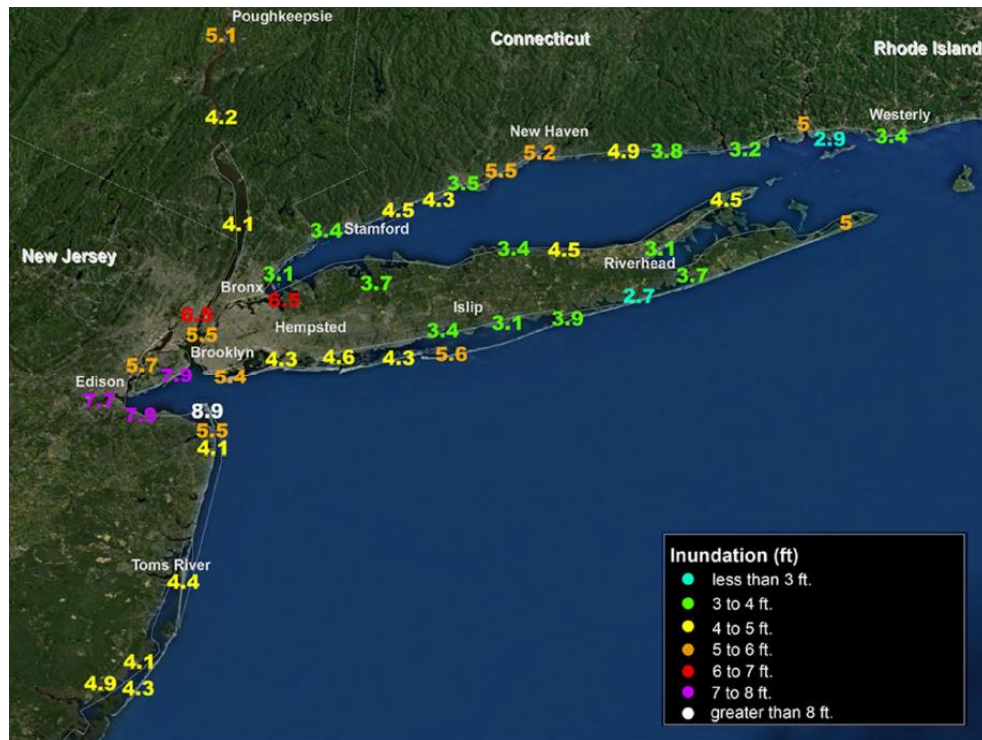


Figure 3. Hurricane Sandy Inundation (Blake E. et al., 2013)

Hurricane Sandy made landfall in the United States as an extratropical cyclone, much weaker than when it hit Cuba days earlier. However, Sandy had a significant spatial impact with winds spanning 945 miles in diameter, making it the largest storm ever observed in the Atlantic (Blake E. et al., 2013). Another important factor for measuring a hurricane is the sustained wind. Numerous weather stations in New York and New Jersey reported sustained winds greater than 70 kts or hurricane strength even though the storm was an extratropical cyclone. The highest recorded wind gust after landfall was 83 kts on the north shore of Long

Island, New York (Blake E. et al., 2013). Rainfall is another impact from hurricanes that can lead to flooding, especially in urban or low-lying areas. The heaviest rain occurred in parts of Maryland, Virginia, and Delaware receiving between five and seven inches. The peak amount of rainfall occurred in Bellevue, Maryland where over 12 inches fell (Blake E. et al., 2013).

The meteorological impact that resulted in the greatest casualties and damage was the storm surge. Sandy caused water levels to rise from Florida to Maine. The highest storm surge and greatest inundation on land occurred in New Jersey and New York, especially in and around New York City (Figure 3). For example, a measurement taken at One World Trade Center in Low Manhattan measured water 4.7 ft above ground level (Blake E. et al., 2013).

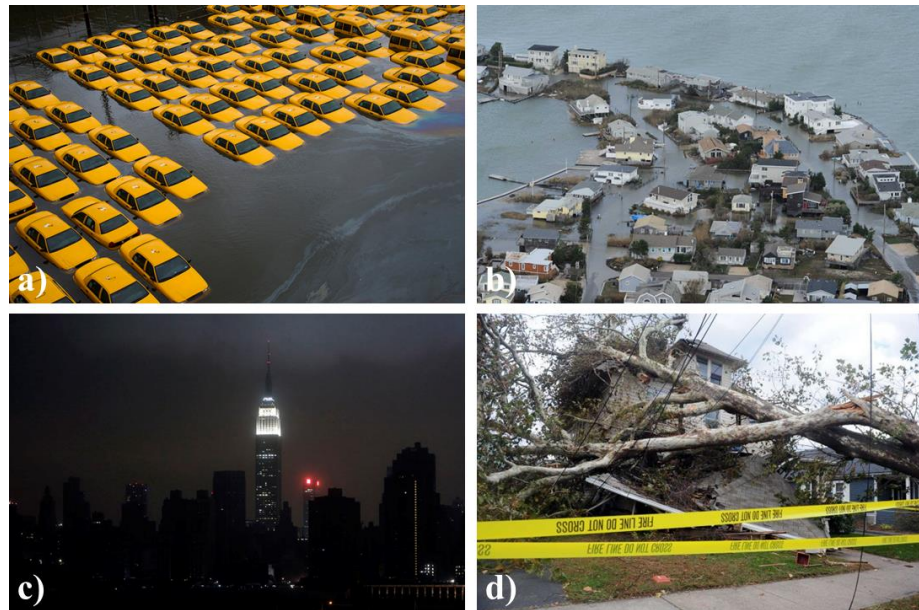


Figure 4. Hurricane Sandy damage in a) Hoboken, NJ¹ b) Westhampton, NY² c) New York City³, and d) Bridgeport, CT⁴

¹ The Associated Press

² Newsday

³ Reuters

⁴ Connecticut Post

Hurricane Sandy would not be categorized as a disaster if it did not directly affect humans. In the United States, unfortunately, Sandy directly caused 72 deaths. The majority of these deaths occurred in New York and because of storm surge. At least 87 deaths were indirectly associated with Sandy in the United States (Blake E. et al., 2013). Most were the result of hypothermia from extended power outages. In terms of infrastructure, over a half million houses were damaged or destroyed and 8.5 million customers lost power at some point during Sandy. In total, Sandy caused about 75 billion dollars in damage, the second-costliest hurricane in United States history (Blake E. et al., 2013).

3.2.2 Data and Data Processing

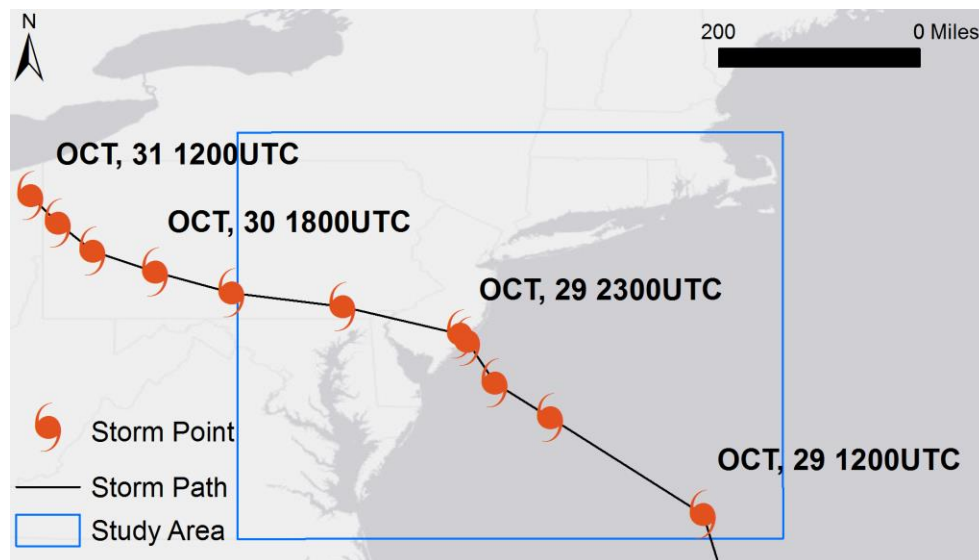


Figure 5. Study area and storm track⁵ of Hurricane Sandy

Before the collection of data, a study area for Hurricane Sandy was selected. Hurricane Sandy affected states in the southeastern United States, like South and North Carolina, as well as many states in the northeast (Figure 2). To capture the disaster beginning right before the stage of landfall, a 400 by 450-mile bounding box was constructed and

⁵ ESRI Basemap

centered at the location of landfall (Figure 5). With a diameter over 800 miles at landfall, the bounding box contains the storm and includes major metropolitan areas, such as New York City and Washington, DC.

For the case study, all data collected were historical in nature. As a result, the data did not stream in real-time into the framework. However, given the appropriate hardware infrastructure, the data ingest streams are feasible.

Social media data was the first to be collected. For the framework to function properly, the social media service or services selected must meet, at minimum, three requirements. First, the data generated on the service should be primarily text-based. In other words, a service like YouTube is not acceptable because the primary data type is video. Other services, such as Flickr or Instagram are acceptable if the image data contains a caption. The second requirement is the service has the option to include the geo-location of the content. The main purpose of the framework is to incorporate location; any service that does not include location would not improve the accuracy of current methods. Finally, the service should be desirable by users to generate disaster event content. A service like LinkedIn is not likely to contain as much disaster related content as Facebook by the nature of their service.

For testing the framework, there are two additional requirements: data access and volume of content generated. Not all social media services allow the public to access their data. Moreover, some services charge a fee for their data. For testing the framework, the service must allow the public to access the data in some form. Additionally, the service must generate enough data to acquire a meaningful sample size. A meaningful size depends on the disaster event, but should be approximately hundreds or thousands of geo-located posts.

Given the requirements for the framework and testing, Twitter is the ideal social media service. Twitter is a text-based microblog with millions of active users. Twitter provides access to its messages or tweets through two public application-programming interfaces (API). The search API allows for retrieval of past tweets based on search criteria (location, keyword, user, etc.). The streaming API retrieves 1% of the most recent tweets based on search criteria.

In total, 12.3 million geo-tagged Tweets were collected from October 28, 2012 – November 7, 2012 (Table 1) using Twitter streaming API. After performing a spatial filter based on the bounding box (Figure 5), the number of Tweets was reduced to 2.8 million.

Table 1. Hurricane Sandy Data

Data Source		Temporal Domain	Spatial Domain	Spatial Resolution	Temporal Resolution	Format
Social network	Tweets	Oct 28 – Nov 7	400 mi x 450 mi	N/A	Milliseconds	Point
Observations	Storm Track	Oct 28 – Oct 31		N/A	Minutes	Point
	Weather Stations			N/A	10-60 Minutes	Point
	Radar			0.5 Degrees x 0.25 km	2-10 Minutes	Raster
Authoritative	Storm Reports			N/A	Minutes	Point
	Watches, Warnings, and Advisories			N/A	Minutes	Polygon
Models	North American Model			12 km	Hourly	Raster
	24 Precip Analysis			4 km	Hourly	Raster

The meteorological data selected for the study offers a variety of ways to measure the hurricane both qualitatively and quantitatively. Additionally, the generated data is from

different sensors, passive, active, and human experts. Different data formats and different spatial and temporal resolutions exist as well (Table 1).

Hurricane track points (Table 1) are produced by the National Hurricane Center. These points indicate the location of the center of the hurricane at important stages in the life of the storm (e.g. change in strength or landfall). By connecting the points, one can get a sense of the overall track of the storm.

The North American Mesoscale Forecast System (NAM) is a high-resolution forecast of hundreds of products. A new NAM model runs every 6 hours predicting the next 84 hours in hourly time steps. Being able to predict where the storm is heading is important for disaster planning purposes. Five NAM products were selected for this study: MSL pressure (for understanding the disaster extent), 1-hour total surface accumulation, surface wind speed, categorical rain (a binary rain classification), and hybrid reflectivity or radar.

In terms of data storage, the radar data was by far the largest dataset at ~45GB due to the temporal resolution (Table 1). The data comes from six different NWS radar stations in the study area. Two products were kept for the rest of the study: base reflectivity (the common weather radar view) and storm total precipitation.

To get more detailed weather measurements on the ground, 128 Automated Weather Observing System (AWOS) stations were used in the study. These stations are primarily located at airports and take measurements at least every hour depending on the conditions. Just like a backyard weather station, AWOS units collect data on many weather variables, such as wind, temperature, dew point, precipitation, and pressure.

The 24-hour precipitation analysis data provides the total precipitation over the last day. Derived from radar and rain gauge reports, this hybrid product provides a high-resolution understanding of how wet it might be in an area (Table 1).

The NWS plays a key role in any meteorological disaster by communicating to both the government and public the severity of the event. The issuing of watches, warnings, and advisories is one well-known way to accomplish this. Table 2 includes a description of the watches, warnings, and advisories collected for Hurricane Sandy.

Table 2. Hurricane Sandy Watch/Warning/Advisory definitions⁶

NWS Issuance	Description
High Wind Watch/Warning	Sustained winds of 40 mph or higher for one hour or more
Small Craft Advisory	Sustained winds of 18 knots to 33 knots or waves of 4 feet or higher
Severe Thunderstorm Watch/Warning	Winds of 58 mph or higher and/or hail 1 inch in diameter or larger
Storm Warning	Sustained winds of 48 knots to 63 knots
Special Marine Warning	Sustained marine convective winds or associated gusts of 34 knots or greater
Hurricane Force Wind Warning	Sustained winds of 64 knots or greater
Gale Warning	Sustained winds of 34 knots to 47 knots
Flood Watch/Warning	Flooding is imminent or occurring
Flash Flood Watch/Warning	Flash flooding is imminent or occurring
Coastal Flood Watch/Warning/Advisory	Moderate to major coastal flooding is occurring or imminent and will pose a serious risk to life and property

Finally, storm reports from the NWS were collected. These reports are a form of volunteer geographic information from experts and the public. Information in the reports includes wind speeds, rain totals, areas of flooding, and damage caused by the storm. The NWS verifies reports through weather data they collected or in person visits. The NWS updates storm reports frequently during a real disaster event.

⁶ NWS - <http://www.weather.gov/lwx/WarningsDefined>

One challenge presented by using the spatial text mining framework is the data variety (Table 1). Before the data transfers into a database, individual processing of the different data sets occurs. For this study, three important standards were established. First, all data would be projected into the WGS84 coordinate reference system. This happens for data analysis and visualization on the web. Second, Shapefiles area for vector data and GeoTIFFs for raster data. This standardization allowed for a simple ingestion into the database. Finally, all date and time parameters were converted into Epoch time. Having the temporal data stored as an integer saves processing time when comparing data.

Each data set presented its own challenges for standardization. For example, the radar data exists natively in a binary format. The Weather and Climate Toolkit⁷ developed by the National Oceanic and Atmospheric Administration (NOAA) read and exported the data as GeoTIFFs. Additionally, the radar and NAM model data originally were stored as single band GeoTIFFs for each product. GDAL⁸ merged the data into multiband GeoTIFFs based on the timestamp. A final challenge involved reading the weather observation data. Each observation comes in a coded text string called a Meteorological Terminal Aviation Routine Weather Report (METAR). Using the Python⁹ package METAR¹⁰, each observation was decoded.

For the purposes of this study, the hurricane track points simulated the event detection phase (Figure 1) of the spatial text mining framework (Table 1). With proper domain

⁷ NOAA Weather and Climate Toolkit - <https://www.ncdc.noaa.gov/wct/>

⁸ GDAL - <http://www.gdal.org/>

⁹ Python - <https://www.python.org/>

¹⁰ METAR by Tom Pollard - <https://pypi.python.org/pypi/metar>

knowledge of the event detection algorithms, one could implement this step in the framework.

3.2.3 Database Selection and Storage

Once the data was prepared, the selection of the databases was made for storing the social media and geographic meteorological data. The variety of social media data poses a challenge for traditional data management following the relational model (Huang & Xu, 2014). Social media services utilize the NoSQL model as a way best manage their social media data. Unlike the traditional relational model, NoSQL implements many different data structures, such as tree, graph, or key-value. The flexibility with NoSQL allows for data from multiple social media services to be stored in one location within the framework.

MongoDB¹¹ was selected as the NoSQL database for this framework. In addition to the reasons state above, MongoDB stores its data in JavaScript Object Notation (JSON) that allows non-uniform fields to be added with no limitations. Most popular programming languages also easily parse JSON. In the event more computing power is required in the framework, MongoDB is scalable allowing multiple servers to store and access the database at a single time. For the meteorological data, PostgreSQL¹² was selected for the framework. PostgreSQL with the PostGIS¹³ extension can store both raster and vector data types, is open source, and provides a wide range of spatial functionality.

¹¹ MongoDB - <https://www.mongodb.com/>

¹² PostgreSQL - <https://www.postgresql.org/>

¹³ PostGIS - <http://postgis.net/>

3.3 Spatial Feature Generation

3.3.1 *Geographic Feature Determination*

The key step in the spatial text mining framework is the spatial feature generation. In this step, the meteorological data is bound to each social media post as a feature through fuzzy logic. Before performing this task, useful spatial features for the hurricane case study should be determined.

Hurricanes are characterized by heavy rain, high winds, and low atmospheric pressure (Roy & Kovordányi, 2012). Thus, it is logical to include these characterizations as spatial features (i.e. rain, wind, and pressure). Since pressure is related to the hurricanes strength, the category of the storm instead of a pressure measurement was used. A few other useful features were also derived from the core geographic features. A flood feature was added because the flooding or storm surge often has the most dangerous impact from a hurricane. Another derived feature added is the presence of an NWS warning meaning an area is in imminent danger of a certain weather hazard. Distance or proximity from the center of the hurricane was the final spatial feature added. Two temporal features were added to add temporal detail to the feature space. One was the date of the social media post. The other was a binary indication of whether the location was currently experiencing the storm or the storm had past.

3.3.2 *Feature Extraction Logic*

The feature extraction algorithm for the Hurricane Sandy case study used fuzzy logic to determine the value associated with each geographic feature. Table 3 details the general conditional logic for each feature. The first decision point in the logic was relation of time to the social media post. If a post happened after the storm dissipated, there was no

meteorological data available. However, this does not necessarily mean a post is not disaster relevant. For example, a person might post a picture of a fallen tree after the storm has passed. To account for this, the rule was to use the time when the storm was closest to the point, but note in another feature that the storm had past. A storm was also denoted as past if the distance exceeded a threshold and the storm was located to the northwest of the post.

With the time sorted out for a social media post, the next step in the logic was to access the data from various sources. Since all data is occurring on different temporal scales, it was highly unlikely that the meteorological data occurred at the same time as the social media post. To solve this problem, each meteorological data product had a valid time criterion. For instance, to use an NWS warning, the post had to have happened within a warning polygon and within the issue and expire times. Another example, to use a storm report, the post must have occurred within 30 minutes of the report.

After the temporal bounds are determined, the social media post must satisfy spatial criteria for each meteorological data product. For raster and polygon data, this involved a simple intersection to attain the attribute value. The numerous point data products required a distance calculation. For example, in addition to the 30-minute time limit, a social media post needed to be within 10 miles of the storm report to attain the attribute value.

The last step in the feature generation algorithm is to rank the data products based on reliability of the data. Each spatial feature had multiple meteorological data source that could explain the feature. For instance, when describing the wind feature, the most accurate data came from weather station observations. Conversely, the weather model provided wind data, but the spatial and temporal resolutions were not as great. In the event a social media post

had values for both data products, the weather station observation was chosen. Table 3 lays out the ranking for each meteorological data product.

Table 3. Spatial Feature Generation Logic

Feature	Description	Data Source Ranking
Disaster Status	Whether or not the storm has past	Twitter and Storm Track
Date	The date of the social media post	Twitter
Distance	The distance from the center of the hurricane	1. Hurricane Track
Storm Category	The category of storm	1. Hurricane Track
Precipitation	How heavy the rainfall is	1. Weather Station 2. Storm Report 3. Radar
Wind	How strong the wind is	1. Weather Station 2. Storm Report 3. NWS Warnings 4. NAM
Flood	Type of flood occurring: flood, coastal, or flash flood	1. Storm Report 2. Weather Station 3. NWS Warnings 4. 24-hr Precipitation Analysis 5. Radar 6. NAM
Warning	NWS warnings	NWS Warnings

3.4 Annotation and Classification

3.4.1 Feature Annotation

After the completion of spatial data processing, all features were annotated with a class for the supervised classification. While some machine learning algorithms can handle numerical data, the algorithms used in the spatial text mining framework rely on categorical data. This approach stayed consistent with the transformation of words into categorical vectors.

Annotating the social media text data requires a degree of domain knowledge.

Previous studies have created different classification schemes to best describe disaster related social media (Gao, Barbier, & Goolsby, 2011; Huang & Xiao, 2015; Imran et al., 2013; Vieweg, Hughes, Starbird, & Palen, 2010; Xiao et al., 2015). One common methodology is to create a two-tiered classification scheme (Imran et al., 2013). First, social media messages are classified as personal, informative, or other. Personal messages are messages only of interest to the author or their immediate circle. Informative messages are of interest to people beyond the author's circle. After the initial filter, informative messages are classified further into five classes, including 1) Caution and Advice (CA), 2) Casualties and Damage (CD), 3) Information Sources (IS), 4) Donation and Aid (DA), and 5) People (Imran et al., 2013; Vieweg et al., 2010). The goal of the two-tiered classification is to describe the overall understanding in disaster events or situational awareness. The classes in Table 4 satisfy the situational awareness domain for a hurricane disaster event.

Table 4. Social media classification scheme (Imran et al., 2013)

Class	Description	Example
Caution and Advice (CA)	Warning or a piece of advice given about a related incident	Flooded neighborhoods in Norfolk and its approaching low tide.
Casualties and Damage (CD)	Information about casualties or infrastructure damage	This tree and power lines are down at the corner of Station Road and Bethlehem Pike in Quakertown.
Information Sources (IS)	A message from an official news source, media or government	@NYCMayorsOffice: Mayor: All @NYCSchools are closed tomorrow.
People	People found or missing	PLEASE RT: IF ANYONE HAS ANY INFO ON THE WHEREABOUTS OF AMANDA LANZONE OF FAR ROCKAWAY, PLEASE PASS IT ON TO @Vicosuave89
Donation and Aid (DA)	Goods or services offered or needed by victims	I don't have any money to donate but I have lots of time, where can I help/volunteer in #Hoboken ? Who do I call ?

Table 5 shows classification scheme of spatial and temporal features, which are not well studied in the literature. The date feature is the date the post was generated. The disaster status feature is an indication of the spatiotemporal relationship of the hurricane and social media post at the time the post was sent (Table 5). For example, if the hurricane is moving away from the post, the feature is classified as past. Conversely, if the hurricane is approaching the post location or is overhead, the classification is present. A distance feature measures the proximity of the post to the center of the storm (Table 5). Distances from the storm were clustered using the DBSCAN clustering algorithm discussed in Section 4.

Table 5. Spatial and temporal feature classification schemes

Feature		Classification Scheme
Temporal	Date	The date of the social media post
Spatiotemporal	Disaster Status	Binary, past or present
Spatial	Distance	DBSCAN clustering algorithm(See Section 4)
	Storm Category	Saffir-Simpson Scale
	Precipitation	Light, moderate, or heavy; NWS Radar Scale ¹⁴
	Wind	Beaufort Scale
	Flood	Flood, coastal flood, or flash flood
	Warning	Binary, warning or no warning

The classification schemes about other spatial and temporal features are less straightforward requiring domain knowledge. In practical applications, domain experts have generated different scales to measure meteorological features. For example, hurricane strength is measured on the Saffir-Simpson Hurricane Wind Scale¹⁵. This scale is represented by the storm category feature. Another useful meteorological scale is the Beaufort Wind Scale¹⁶. The scale provides both land and sea descriptions for different strengths of wind from clam to hurricane force. The wind feature is calculated from the

¹⁴ NWS - http://www.weather.gov/media/publications/front/06nov_Front.pdf

¹⁵ <http://www.nhc.noaa.gov>

¹⁶ NWS - <https://www.weather.gov/tbw/beaufort>

Beaufort Wind Scale. Precipitation intensity feature is calculated using the NWS Radar Scale¹⁶, which converts radar reflectivity into a precipitation intensity. The final two meteorological features flood and warning, utilize the NWS categorical watches, warnings, and advisories classification. For example, if a post occurred in a flash flood warning, flash flood would be assigned to the flood feature and the warning feature set to true.

3.4.2 *Feature Classification*

Where the spatial feature generation is the key component of the framework, the text mining component is necessary for generating the desired outcome. The goal of this component is to determine if a social media post is disaster relevant or not. Accuracy is the primary indicator of assessment and relates directly to research question two. With the text mining features prepared by the spatial feature generation component, this step required an appropriate choice for text mining algorithms to establish the classification model. Right now, there is a variety of classification algorithms available, such as K-nearest neighbor, decision trees, logistic regression, and neural networks. However, for efficient classification, the Naïve Bayes and support-vector machines models are commonly used (Ashktorab et al., 2014; Huang & Xiao, 2015; Spinsanti & Ostermann, 2013; Takahashi, Tandoc, & Carmichael, 2015). Both methods are supervised classifiers requiring training data to function properly.

To generate a training sample for classification, the 2.8 million social media posts were filter using the hashtag #sandy. With 33,963 posts remaining, a random sample of 5,000 posts was taken. The author primarily accomplished manual classification (Table 4) of the sample. When the class definition of a post was not obvious, experts gave feedback to finalize the class.

3.5 Spatiotemporal Distribution of Sample Data

Before running the classification experiment (Section 4), understanding the spatiotemporal distribution of the labeled sample data was important for interpreting the results. In order to ensure consistency and accuracy, the author manually labeled all sample data twice. Any discrepancies in the classification between the two rounds of labelling were resolved with experts. Of the 5,000 social media posts sampled, 1,919 posts were labeled as informative.

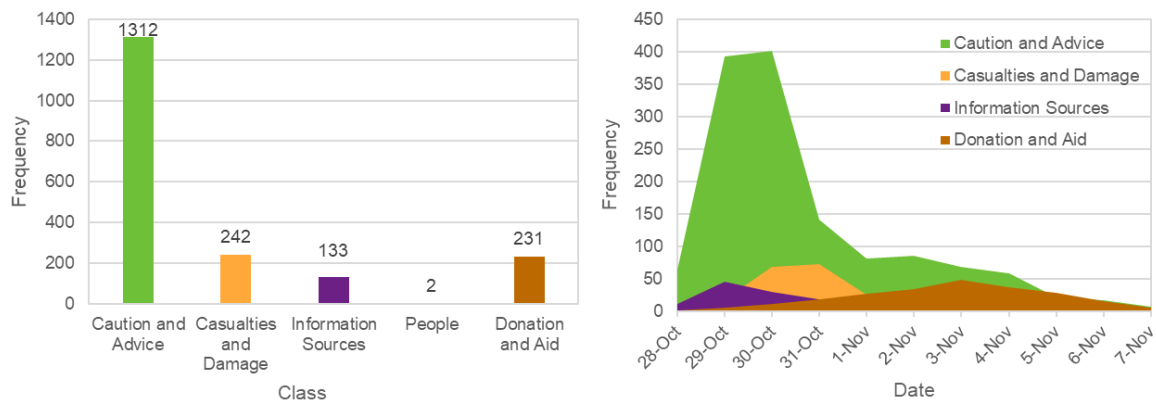


Figure 6. The frequency (Left) and temporal distribution (Right) of informative posts

Shown as Figure 6, the breakdown in posts per class were not surprising. Caution and advice was the most general class and contained most samples. The CA posts include disaster preparation information about food, shelter, gas, and other resources. During the disaster, the CA class contained posts with information about inaccessible areas or lack of resources. Additional posts about closures and traffic information fell into this class. CD and DA classes had roughly the same number of messages. Both classes occur during and after the disaster event. The IS class had very few sample without a clear temporal pattern. Analysis of the IS class is contained in section 4 of the thesis. People missing or found class, had the lowest amount of posts in the sample. This class was removed from the classification

experiments due to the low sample count. Figure 7 shows the spatial distribution of the data. Most posts are clustered in the three largest metropolitan areas: Philadelphia, Washington, and New York. Another area worth noting is along the New Jersey coast where the most severe damage was located.

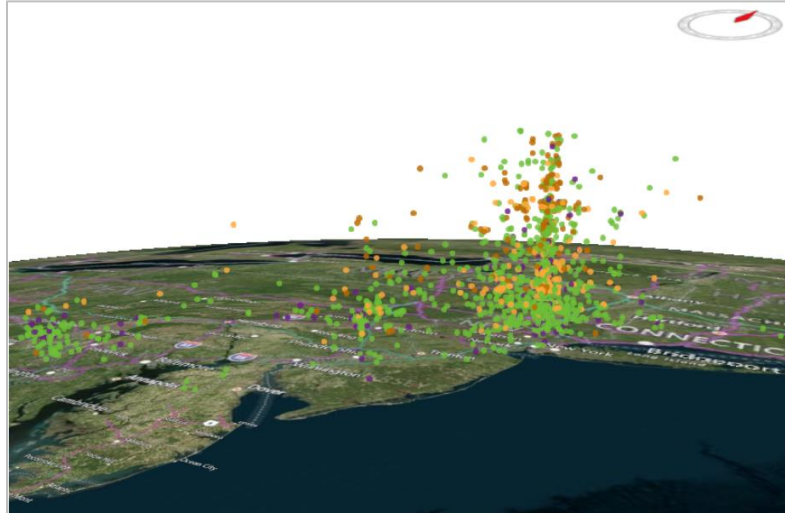


Figure 7. Spatial and temporal distribution of informative posts. Post time is displayed in the vertical axis beginning with October 28, 2012 at the bottom.¹⁸

The last step in the spatial text mining framework for disaster events is the use of the results for disaster applications. A possible application of the framework is a front-end geoanalytics tool. Viewing all the datasets in a dynamic visualization allows disaster managers and planners to gain insights from the data to make decisions. The data was uploaded into RealEarth¹⁹, a geovisualization tool developed at the Space Science and Engineering Center at the University of Wisconsin. Users can view data at individual time steps or animate through the event.

¹⁸ NASA World Wind

¹⁹ RealEarth™ - <https://realearth.ssec.wisc.edu/>

4. Experimental Design and Analysis

The remainder of this study focuses on the evaluation and validation of the spatial text mining framework (Figure 1), specifically the classification step to answer the two research questions laid out in the beginning of this thesis. First, which of the spatiotemporal features for the Hurricane Sandy case study are key to describing and then classifying disaster relevant social media posts? Second, can the key features improve accuracy of the current text-only classification approach? This section first describes the experimental design based on these research questions, followed by the results from the spatiotemporal feature determination. Finally, the section discusses the results of the feature combination experiments.

4.1 Experimental Design

4.1.1 Selected Experiments

Table 6. Spatial text mining framework experiments

Feature Type	Feature	Text-only	All	Proximity	Proximity and Time	Enhanced Proximity and Time	Meteorological
Text	Text, 1-gram	X	X	X	X	X	X
	URL		X			X	
Temporal	Date		X		X	X	
Spatio-temporal	Disaster status		X			X	
Spatial	Distance		X	X	X	X	
	Storm Category		X				X
	Precipitation		X				X
	Wind		X				X
	Flood		X				X
	Warning		X				X

To determine if geographic features improve the accuracy of current text-only methods, a series of experiments were derived based on hypotheses of the results (Table 6).

A text-only experiment would serve as the control to the other spatial experiments. Since the features included were determined to describe the case study the best from the literature, the next experiment used all features. However, utilizing all features poses two problems. First, using too many features can generate noise in the model leading to poor results. Second, the goal of this research is to identify the key spatial features or the minimal number of features that give the best accuracy.

Four additional experiments tested the number of features based on hypotheses derived from domain knowledge. In a disaster event, distance from the center of the event is important for determining the relevancy of social media posts (de Albuquerque et al., 2015). Thus, the hypothesis is that distance by itself would improve the classification accuracy when combined with text. Another feature hypothesized to aid in the classification accuracy was the date feature. The classification scheme utilized in this and many other research studies is by nature time sensitive. For example, social media posts about donation and aid are more likely to occur during and after a disaster event (Figure 6). Since the machine learning algorithms selected for this study are dependent on probabilities, the expectation is that the date feature will improve the classification. Because of both hypotheses, the expectation is combining the distance and date feature with the text feature would give the best results. Out of curiosity, a third experiment related to distance and time was created which combined what the author determined to be the best features for describing the study based on empirical tests of the features. The last experiment combined the remaining spatial features related to the meteorological data.

4.1.2 Performance Measures

To assess the results of different classification experiments, the study incorporated a set of performance measures common with model assessment. It is worth noting again the methodology used in this study. The author labeled a random sample of 5,000 social media posts into informative and uninformative classes. Then the author labeled the remaining 1,920 informative posts into the disaster relevant classes (Table 4).

Running the labeled data through a supervised classifier involves taking a subset of the data to train the classifier and the remainder of the labeled data to test the classifier. One potential error associated with this approach is for bias to exist in the training data resulting in suboptimal results. Cross-validation is a technique used to evaluate the model by partitioning the sample data into k equally sized subsamples. From there $k-1$ subsamples represent training data with a single subsample as the test data. The cross-validation process repeats k times with the results averaged from the k iterations. For this study, the value of k is ten. Moreover, a stratified version of cross-validation ensures a proportional representation of each class in each subsample.

After the cross-validation is complete, several metrics determine the classifiers' accuracy: precision, recall, and F-1 score. The precision of a class is the number of correctly classified samples divided by the total number of samples classified as a class. In other words, precision is a measure of relevancy. On the other hand, recall is the number of correctly classified samples divided by the actual number of samples for a class. Another way to think about recall is as a measure of sensitivity in the classification. In theory, precision and recall are unrelated, but in practice, high precision usually leads to a lower recall and vice versa. To overcome this problem of metrics, this study also used F-score to

evaluate the results. The F-score is a single measure based on the harmonic mean of the precision and recall. Overall accuracy for an experiment was determined by averaging the F-score over the different classes.

4.1.3 Classifier Selection

Before the primary feature experiments, a classification experiment was conducted to determine the most accurate classification algorithm using only the text feature. The remainder of experiments used the most accurate algorithm from this experiment. To simplify the testing process a single Python package called scikit-learn²⁰ was used for all experiments. Scikit-learn is a powerful and popular open-source machine-learning library containing hundreds of useful algorithms from clustering to regression. What makes scikit-learn so powerful is its interoperability with other Python scientific libraries allowing for simple interchangeability of algorithms. From the available supervised classification algorithms, three were selected based on previous text mining studies: Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM). In all three evaluation metrics, the SVM had higher accuracy values than the other two algorithms (Table 7). The reason for these results is relatively straightforward. One of the assumptions of Naïve Bayes is conditional independence, but given a large feature vector space for the text, it is difficult to form discrete classes. Moreover, Naïve Bayes and Logistic Regression have difficulty fitting the model to uneven class distributions. The remaining experiments utilized the SVM algorithm.

Table 7. Text-only algorithm experiment

Algorithm	Avg. Precision	Avg. Recall	Avg. F-score
Naïve Bayes	0.809	0.759	0.694
Logistic Regression	0.800	0.782	0.745
Support Vector Machine	0.8187	0.821	0.806

²⁰ scikit-learn - <http://scikit-learn.org/stable/>

4.2 Feature Determination

4.2.1 Spatial Features

As discussed in Section 3.5.1, determining the feature annotation for the distance from the center of the disaster was not a trivial task. Originally, the distance feature contained classes for different distance ranges loosely based on the radius of the storm. For example, if a user posted more than 250 miles away from the storm, it was classified as irrelevant because the post was too far away from the storm. While this simple feature annotation provided reasonable results in the classification, there are several issues using this strategy. First, having fixed distance ranges does not account for the potential temporal variability associated with the different text classes. For instance, a social media post classified as DA is more likely to occur after the storm has passed. However, the distance from the center of the storm might be greater than the 250-mile threshold. Second, an arbitrary distance classification does not account for the spatial patterns that exist in the sample data. Figure 7 shows the sample data clusters around metropolitan areas. In theory, the social media posts nearest to one another have similar experiences from the disaster.

To test the impacts of the distance feature annotation, an experiment was conducted using DBSCAN to cluster the posts based on distance from the center, the date, and the latitude and longitude. DBSCAN is a density-clustering algorithm which clusters points based on a minimum number of samples within a specified epsilon or search distance (Ester, M. et al., 1996). High-density points cluster into groups and low-density points classify as noise. Table 8 highlights the results of the experiment when the clusters are used with text to classify the social media posts. The minimum number of samples remained constant for each

test and the epsilon was changed until a consistent number of clusters were found. Overall, the results of the clustering were as expected. The highest F-score occurs when distance and date are the clustering features (Table 8). However, using the date feature in the DBSCAN and as a feature in the classification (Table 6) could result in feature redundancy. To minimize feature redundancy, the second-best trial, latitude, longitude, and distance, was selected as the classification of the distance feature (Table 8).

Table 8. DBSCAN analysis for distance feature with minPts as 4 and different eps values

DBSCAN Features	Eps	Clusters	Avg. Precision	Avg. Recall	Avg. F-score
Distance	0.08	5	0.812	0.817	0.801
Distance & Date	0.08	7	0.824	0.827	0.813
Lat/Lon & Date	0.4	6	0.815	0.820	0.804
Lat/Lon & Distance	0.4	7	0.821	0.823	0.807
Lat/Lon	0.2	9	0.819	0.821	0.805

4.2.2 Temporal Feature

In addition to the distance feature, the temporal feature was hypothesized to be important for improving the text classification. As previously mentioned, the text classification scheme (Table 4) is associated with time. Before performing the experiments, the social media data was visually inspected based on time (Figure 8). A day before the disaster event, the few messages posted primarily related to the CA class (Figure 8a). This type of result meets the expectation of a disaster because no damage exists prior to the event.

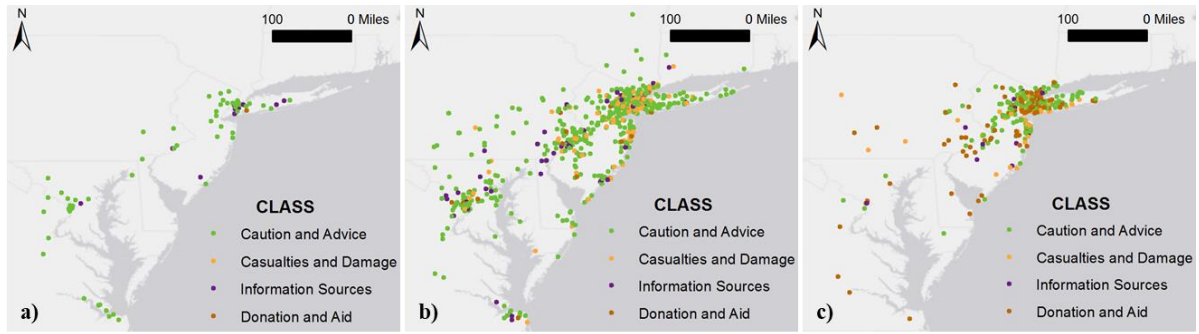


Figure 8. Spatial distribution of sample data by different classes at different disaster stages, including before (Oct. 28; a), during (Oct. 29-31; b), and after disaster event (Nov. 1-7; c).

Spatially, the posts before the storm are in metropolitan areas (**Figure 8a**). As the storm affects the disaster area, the IS and CD classes become more prominent (**Figure 8b**). While the frequency of posts increases in the metropolitan areas, the areas hit hardest along the New Jersey coast extending inland become flooded with messages. Finally, after the storm dissipated the prominent class for social media posts is DA (**Figure 8c**). Major posts after disaster are in significantly impacted areas by the storm.

4.3 Feature Combination Performance

4.3.1 Combination Performance

The goal of the feature combination experiment is to answer the two research questions for this study (Table 6). First, determine which feature or features best describes the disaster in the classification models. The hypothesis is distance from the center and time will play the largest role. Second, does the addition of features improve the accuracy of the classification compared to the current text-only approach? Based on the evidence of natural spatiotemporal clustering of posts, the hypothesis is accuracy improves. Using an SVM with 10-fold cross-validation, each experiment was performed (Table 9).

Table 9. Feature combination experiment results

Experiment	Class	Precision	Recall	F-score
Text-only	CA	0.821	0.960	0.885
	CD	0.732	0.417	0.532
	IS	0.786	0.496	0.608
	DA	0.914	0.645	0.756
	Averages	0.819	0.821	0.806
All	CA	0.844	0.950	0.894
	CD	0.700	0.529	0.602
	IS	0.800	0.391	0.525
	DA	0.851	0.714	0.777
	Averages	0.824	0.830	0.817
Proximity	CA	0.830	0.954	0.888
	CD	0.736	0.471	0.574
	IS	0.753	0.459	0.570
	DA	0.879	0.662	0.756
	Averages	0.819	0.824	0.810
Proximity & Time	CA	0.841	0.949	0.892
	CD	0.719	0.508	0.596
	IS	0.800	0.421	0.552
	DA	0.848	0.723	0.780
	Averages	0.824	0.830	0.818
Enhanced Proximity & Time	CA	0.847	0.947	0.894
	CD	0.706	0.546	0.615
	IS	0.814	0.429	0.562
	DA	0.846	0.714	0.775
	Averages	0.827	0.832	0.822
Meteorological	CA	0.826	0.959	0.888
	CD	0.745	0.459	0.568
	IS	0.771	0.406	0.532
	DA	0.898	0.684	0.776
	Averages	0.821	0.824	0.809

Each experiment, which combined spatial and temporal feature with text, improved the overall accuracy compared to the text-only classification (Table 9). The enhanced proximity and time experiment had the highest average recall (0.832) and F-score (0.822) of the experiments. Adding only meteorological features had the least amount of improvement to the recall and F-score (Table 9). The results from these experiments indicate time and

proximity to the disaster are the best features to describe the disaster event. Moreover, combining these features with text does improve the overall accuracy of the classification.

Assessing individual classes reveals a few patterns that exist in the different experiments. First, the caution and advice category made up 68% of the total sample and performed relatively the same for each experiment (Table 9). Thus, adding spatial features had minor impact on the class compared to text-only. The result is likely due to the broad class spatial distribution within the sample (Figure 7) and the high number of samples compared to the other classes (Figure 6). Second, in general the damage and donation classes improved dramatically in the recall resulting in higher F-scores (Table 9). For disaster relevant information retrieval, identifying a greater number of disaster relevant posts accurately or increasing the recall is important. Moreover, the combination of time and distance features improves recall the best, which makes sense given the class definitions and sensitivity to space and time. Finally, the IS class stayed consistent in precision among the experiments, but decreased dramatically in recall performance with the addition of spatial and temporal features (Table 9). This class was the least represented in the sample data used for the experiments. Furthermore, it does not have a strong spatial relationship. Finally, there is likely confusion in the classification of the text because of class overlap. For example, if a social media post mentions the mayor giving a damage report, this post should be considered an information source because it is a secondary source. Unfortunately, the classifier could mistake the text as related to the CD class.

4.3.2 Sensitivity Assessment

To form a better understanding of the classification of social media posts, a sensitivity assessment was performed by comparing the classification performance using the text-only and enhanced proximity and time models. The results were binned into five categories for each class. If the posts were classified correctly in both experiments, they were classified as correct and vice versa for the wrong posts. These posts are not sensitive to the addition of spatial and temporal features. The remaining categories, correct to wrong, wrong to correct, and wrong to different wrong class, indicate a change in classification based on the addition of spatial and temporal features.

Table 10. Sensitivity analysis for the enhanced proximity and time experiment by class

Class	Correct	Wrong	Correct to Wrong	Wrong to Correct	Wrong to Diff. Wrong
CA	82.75%	13.30%	1.77%	1.98%	0.20%
CD	45.99%	7.49%	18.72%	24.60%	3.20%
IS	77.14%	14.28%	4.29%	4.29%	0%
DA	73.85%	5.13%	7.18%	10.77%	3.07%

The results of the sensitivity analysis support the justification of the performance results from the classification experiment (Table 10). Generally, more posts shifted from wrong to correct than correct to wrong when classified with spatial and temporal features. However, for the CA and IS classes, these values stayed relatively the same indicating little sensitivity. On the other hand, the CD and DA classes had larger differences in the between the categories which supports the increase in recall for both classes when spatial and temporal features were added. The spatial autocorrelation is positive when looking at the sensitivity categories. Clustering of sensitivity posts by category suggests an influence of spatial features within the classification.

In summary, the addition of spatial and temporal features to the text classification increases the overall classification accuracy compared to text-only classification. For the hurricane case study, distance from the storm and time of the post are the most influential features for describing the disaster event based on the classification results. The performance impact on different classes, however, varies. Classes defined without a specific spatial or temporal context performed the same or worse with the added features.

5. Conclusion and Future Study

5.1 Conclusion

This thesis presents an enhanced text mining framework that incorporates spatial and temporal data into the classification of disaster-relevant social media posts. With the popularity of social media higher than ever before, disaster management represents one application well-positioned for leveraging the data. Using citizens as sensors allows disaster managers and planners to understand the situational awareness during a disaster event.

To filter through the high volume of social media data generated to find disaster relevant posts, current approaches use machine learning. These approaches focus on text mining keywords to classify disaster relevant posts. However, the text mining approaches are not without error. Given the high degree of variability in natural language processing, current methods have difficulty understanding context. Additionally, data rich multimedia presents a data source unreachable via current methods.

The goal of the spatial text mining framework is to improve the current text-only methods by improving the classification accuracy. The framework ingests and processes spatial data that can describe the disaster at a given place in time. In doing so, the spatial and temporal data enhances the text data by providing additional situational awareness. Then the framework combines the spatial and temporal data in the form of vector features with the text data and classifies the posts based on relevance using a supervised classification algorithm. The last step in the framework is the visualization of classified disaster relevant social media data for disaster managers and planners to utilize for decision making.

Data from Hurricane Sandy provided a means for testing the framework. The disaster presents several big data challenges in data volume and variety. Collection, processing, and

standardization varied for each meteorological data set. Using fuzzy logic, spatial features (e.g. distance, wind, flood, precipitation...etc.) were bound to each social media post in the sample data set. Ahead of the supervised classification, the author and experts performed manual labelling of each post into disaster relevant classes. Numerous experiments tested the performance of the spatial features with the text feature to answer the two research questions.

Based on the results, the distance from the disaster and the position relative to the post were determined to be the key features in addition to text for describing the disaster event. Moreover, the addition of spatial and temporal features to the text classification did improve the overall classification accuracy compared to current methods. From the experiments, the impact of spatial features on individual classes was relative to the class definition. The additional features had less of an impact on classes (e.g. caution and advice or information sources) with a greater spatial or temporal distribution than classes (e.g. casualties and damage or donation and aid) whose class definitions involved spatial and temporal components. Social media is not going away anytime soon. For disaster applications related to social media data, this study indicates the need for the incorporation of geographic data sources to improve data retrieval and provide greater situation awareness.

5.2 Future Direction

With limited time and resources for this study, this topic requires additional effort for improvement to the framework. Four limitations and potential future directions are:

Feature definitions: Currently, research of spatial feature creation and definition is nonexistent. For this study, when possible authoritative definitions from government sources defined the feature bins. The development of the model is only as good as the definitions of

the features. More investigation into optimal spatial feature creation is a possible step to improving the classification results.

Classification algorithms: This study focuses on the common machine-learning algorithms for classifying text data. However, given the hybrid nature of the data going into the classifier, the assessment of other algorithms is important for finding the preferred framework classifier. One possible solution is to utilize an ensemble method that combine the predictions of several estimators for a given algorithm. Moreover, based on the sensitivity of different classes to distinctive features, hierarchical algorithms like decision trees are a solution to handle the complicated class logic.

Different disaster events: The original intent of this study was to assess the differences in the framework results across disaster events. One of the challenges to completing this task is the high degree of domain specificity required for each disaster. Fortunately, the framework can handle this sort of variation between disasters.

Classification schema: The results of this study show how the addition of spatial features influences classes of disaster relevant social media data. As discussed in the literature review, many different classification schema exist for defining disaster events. Additional research is necessary to understand and better define a classification schema that can leverage the spatial features to extract relevant data.

References

- Acar, A., & Muraki, Y. (2011). Twitter for crisis communication; lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*.
- Alberts, T., Chilson, P., Cheong, L., & Palmer, R. (2011). Evaluation of Weather Radar with Pulse Compression: Performance of a Fuzzy Logic Tornado Detection Algorithm. *Journal of Atmospheric Oceanic Technology*.
- Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*, (May), 354–358. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-84905845531&partnerID=40&md5=ee57e6c3d9498b083428cdac67d83396>
- Bakillah, M., Li, R.-Y., & Liang, S. H. L. (2014). Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan. *International Journal of Geographical Information Science*, 8816(January 2016), 1–22. <https://doi.org/10.1080/13658816.2014.964247>
- Blake, Eric S.; Kimberlain, Todd B.; Berg, Robert J.; Cangialosi, John P.; Beven, J. L. . (2013). *Tropical Cyclone Report: Hurricane Sandy*. National Hurricane Center. Retrieved from http://www.nhc.noaa.gov/data/tcr/AL182012_Sandy.pdf
- Bruns, A., & Liang, Y. E. (2013). Tools and methods for capturing Twitter data during natural disasters. *First Monday*, 17(4–2), 1–17.
- Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, 17(1), 124–147. <https://doi.org/10.1111/j.1467-9671.2012.01359.x>

- de Albuquerque, J. P., Herfort, B., Brenning, A., & Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 8816(June), 1–23. <https://doi.org/10.1080/13658816.2014.996567>
- De Longueville, B., Annoni, A., Schade, S., Ostlaender, N., & Whitmore, C. (2010). Digital Earth's Nervous System for crisis events: real-time Sensor Web Enablement of Volunteered Geographic Information. *International Journal of Digital Earth*, 3(3), 242–259. <https://doi.org/10.1080/17538947.2010.484869>
- Ester, M., H. P. Kriegel, J. Sander, and X. X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press*, 2, 226–231. <https://doi.org/10.1016/B978-044452701-1.00067-3>
- Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3), 10–14. <https://doi.org/10.1109/MIS.2011.52>
- Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3), 231–241. <https://doi.org/10.1080/17538941003759255>
- Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1, 110–120. <https://doi.org/10.1016/j.spasta.2012.03.002>
- Ford, R. (2011). Earthquake: Twitter users learned of tremors seconds before feeling them. *The Hollywood Reporter*. Retrieved from <http://www.hollywoodreporter.com/news/earthquake-twitter-users-learned-tremors->

226481

- Fuchs, G., Andrienko, N., Andrienko, G., Bothe, S., & Stange, H. (2013). Tracing the German centennial flood in the stream of tweets: first lessons learned. *In Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, Orlando, FL, USA, 31-38.
- Halteren, H. van, Zavrel, J., & Daelemans, W. (2001). Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics*, 27(2), 199–229. <https://doi.org/10.1162/089120101750300508>
- Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R., ... Griffith, S. a. (2014). Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*, 39(1), 1–22. <https://doi.org/10.1111/disa.12092>
- Huang, Q., Cervone, G., Jing, D., & Chang, C. (2015). DisasterMapper: A CyberGIS framework for disaster management using social media data.
- Huang, Q., & Xiao, Y. (2015). Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery. *ISPRS International Journal of Geo-Information*, 4(3), 1549–1568. <https://doi.org/10.3390/ijgi4031549>
- Huang, Q., & Xu, C. (2014). A data-driven framework for archiving and exploring social media data. *Annals of GIS*, 20(4), 265–277. <https://doi.org/10.1080/19475683.2014.942697>
- Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Practical extraction of disaster-relevant information from social media. *Proceedings of the 22nd International*

Conference on World Wide Web, 1021–1024. <https://doi.org/10.1145/0000000.0000000>

Jain, S. (2015). Real-Time Social Network Data Mining For Predicting The Path For A Disaster. *Computer Science Theses*. Retrieved from http://scholarworks.gsu.edu/cs_theses/79%5Cnhttp://scholarworks.gsu.edu/cs_theses/79/

Keim, M. E., & Noji, E. (2011). Emergent use of social media: a new age of opportunity for disaster resilience. *Am J Disaster Med*, 6(1), 47–54.

Lakshmanan, V., & Smith, T. (2009). Data mining storm attributes from spatial grids. *Journal of Atmospheric and Oceanic Technology*, 26(11), 2353–2365.
<https://doi.org/10.1175/2009JTECHA1257.1>

Landwehr, P. (2014). Social Media in Disaster Relief: Usage Patterns, Data Mining Tools, and Current Research Directions. *Data Mining and Knowledge Discovery for Big Data*, 225–257.

Lindsay, B.R. (2011). *Social media and disasters: current uses, future options, and policy considerations, congressional research service* (Report No. R41987)

Mills, A., Chen, R., Lee, J., & Rao, H. R. (2009). Web 2.0 emergency applications: how useful can Twitter be for emergency response? *Journal of Information Privacy and Security*, 5(3), 3–26.

Roy, C., & Kovordányi, R. (2012). Tropical cyclone track forecasting techniques - A review. *Atmospheric Research*, 104–105, 40–69. <https://doi.org/10.1016/j.atmosres.2011.09.012>

Schnebele, E., & Cervone, G. (2013). Improving remote sensing flood assessment using volunteered geographical data. *Natural Hazards and Earth System Science*, 13(3), 669–677. <https://doi.org/10.5194/nhess-13-669-2013>

- Schnebele, E. K., Cervone, G., & Waters, N. (2014). Road assessment after flood events using non-authoritative data. *Natural Hazards and Earth System Science*, 14, 1007–1015. <https://doi.org/10.5194/nhess-14-1007-2014>
- Shklovski, I., Burke, M., Kiesler, S., & Kraut, R. (2010). Technology Adoption and Use in the Aftermath of Hurricane Katrina in New Orleans. *American Behavioral Scientist*, 53(8), 1228–1246. <https://doi.org/10.1177/0002764209356252>
- Spinsanti, L., & Ostermann, F. (2013). Automated geographic context analysis for volunteered information. *Applied Geography*, 43, 36–44. <https://doi.org/10.1016/j.apgeog.2013.05.005>
- Sutton J., Palen L., & Shklovski I. (2008). Backchannels on the front lines: emergent uses of social media in the 2007 southern California wildfires. *5th international ISCRAM conference, Washington, DC*.
- Takahashi, B., Tandoc, E. C., & Carmichael, C. (2015). Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines. *Computers in Human Behavior*, 50, 392–398. <http://doi.org/10.1016/j.chb.2015.04.020>
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–240.
- Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, 1079. <https://doi.org/10.1145/1753326.1753486>
- Xiao, Y., Huang, Q., & Wu, K. (2015). Understanding social media data for disaster management. *Natural Hazards*, 79(3), 1663–1679. <https://doi.org/10.1007/s11069-015->

1918-0

Zlatanova, S., Zlatanova, S., Holweg, D., & Holweg, D. (2004). 3D Geo-Information in Emergency Responce: a Framework 1. *Proceedings of the Fourth International Symposium on Mobile Mapping Technology (MMT'2004), March, 29–31*. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:3d+geo-information+in+emergency+responce:+a+framework+1#0>