Geographic Context-aware Text Mining: Enhance Social Media Message Classification for Situational Awareness by Integrating Spatial and Temporal Features

Christopher Scheele[1], Manzhu Yu[2*], Qunying Huang[1*]

[1]Spatial Computing and Data Mining Lab, Department of Geography,
University of Wisconsin-Madison, Madison, Wisconsin, 53706, {cjscheele,
qhuang46}@wisc.edu
[2]Department of Geography,
Pennsylvania State University, University Park, PA 16802, mqy5198@psu.edu

*Corresponding author: Manzhu Yu, mqy5198@psu.edu; Qunying Huang,
qhuang46@wisc.edu

Abstract: Social media data are widely used in disaster management for event detection, response, and recovery. To find disaster relevant social media messages and automatically categorize them into different classes (e.g. damage or donation), current approaches utilize natural language processing 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 and ignoring the geographic context of the messages when posted. Meanwhile, a disaster relevant social media message is highly sensitive to its posting location and time. Thus, additional features related to space and time could be useful for differentiating relevant posts by informing its geographic context, and therefore improving purely text-based approaches. However, limited studies exist to explore what spatial features and the extent of how temporal, and especially spatial features can aid text classification. To fill the research gap, this paper proposes a context-aware text mining method to incorporate spatial and temporal information derived from social media and authoritative datasets (e.g., Earth observations, physical model output, official reports), along with the text information, for classifying disaster relevant social media posts. With the 2012 Hurricane Sandy as a case study, we designed and demonstrated how diverse types of spatial features, such as wind, flooding, and proximity, and temporal features can be derived from spatial data, and then used to enhance text mining. The deep learning based method, convolution neural networks, and commonly used machine learning algorithms (e.g., support vector machine), assessed the accuracy of the enhanced text-mining method. The performance results of different classification models generated by various combinations of textual, spatial, and temporal features indicate that additional spatial and temporal features help improve the overall accuracy of the classification by 4 percentage points on average. This study demonstrates the need and provides a guidance for the incorporation of geographic data sources to improve data retrieval while leveraging social media for disaster applications.

Keywords: spatial data science; spatially enabled text mining; situational awareness; deep learning; GeoAI; spatial features

1. Introduction

The rise in social media such as Facebook, Twitter, Flickr, LinkedIn, YouTube, and many others, in the past decade has changed the way people interact with each other, and gain situational awareness (SA) during a disaster event. 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. 2015). Disaster managers, on the other hand, can also gather social media data to monitor disaster events in real-time. With this approach, citizens involved with the disaster act as "sensors" providing geo-located information to supplement authoritative data sources (De Longueville, Smith, and Luraschi 2009). Having a geographical SA through social media enables the identification of areas with infrastructure damage, affected people, and evacuation zones (Huang and 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 (Ashktorab et al. 2014; de Albuquerque et al. 2015; Huang et al. 2015). One of the typical approaches, known as the text-based approach, involves text-matching by searching for specific keywords or groupings of words using machine learning algorithms to determine if the social media data is relevant (Ashktorab et al. 2014; Huang and Xiao 2015; Bakillah, Li, and Liang 2015; Landwehr and Carley 2014). This approach builds a classifer that categorizes text on the existence of keywords. Although text-based classification is a fast way to organize large datasets, it has two major limitations while identifying or differentiating all social media data related to a disaster:

- Uncertainty and variability in language used on social media (Bruns and Liang 2012). 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 and Liang 2012).
- Social media data can be easily misidentified by purely relying on text without considering the geographic context. For example, during a flooding event, the message "the water is very high right now" could have different contexts. While an individual on a riverfront within the disaster area most likely indicates, "the water level is high", another individual at home far from the impact region could mean, "the water in the bathtub is high."

To overcome these limitations, this study develops a context-aware texting mining method that integrates spatial and temporal information, revealing the geographic context of a social media post, to classify whether a tweet is relevant or not to a disaster. A variety of spatial data could contribute to inform the geographic and environmental context. Some spatial data are highly domain-specific. For example, during a flood event, information about how much rain has fallen and areas of lower elevation are 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 (e.g., humidity) in the area or wind patterns are more relevant data. Other spatial data are less domain specific, the most prominent being proximity to the disaster. Based on Tobler's First Law of Geography,

individuals posting on social media about a tornado going through their town are more related to the disaster than an individual 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 and Xiao 2015; Zou et al. 2018). Therefore, the first objective of this paper is to identify spatial data from which spatial and temporal features can be produced to help inform the geographic context of text messages.

In addition, current text-based classification models cannot directly assimilate raw spatial (e.g., remote sensing imagery) and temporal 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 well studied, but vital for this research. Therefore, this paper also aims to examine how spatial and temporal features can be derived, and to provide a reference on the utilization of the spatial and temporal data as features in the classification models. Furthermore, existing studies show that deep learning based methods, such as convolution neural network (CNN) and recurrent neural network (RNN), significantly outperformed traditional machine learning approaches, such as support vector machine (SVM; Joachims 1998) for text classification tasks (Yu et al. 2019). However, it is not clear whether the integration of spatial and temporal features will improve the deep learning based methods. As such, this study will then compare the performance of CNN models, one of the popular deep learning based methods, with traditional text-mining approaches (e.g., SVM). In addition, we will further identify the spatial and temporal features that are useful to improve the accuracy of CNN models based on its performance evaluation while combining varying spatial, temporal and textual features.

To sum up, the following contributions are addressed in this research:

1. First, the paper introduces a methodology for integrating geographic context into classifying disaster relevant social media datasets by fusing spatial data with social media. The method addresses the shortcomings of utilizing only text to identify and extract disaster relevant social media data when considering geographic context is necessary.
2. Second, this paper demonstrates how to best process spatial and temporal data to derive associated features for classifying and identifying disaster relevant information.
3. Third, this paper 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.
4. Finally, this work evaluates both traditional machine learning algorithms (e.g., SVM), and the state-of-the-art work deep learning based method, CNNs, on SA information classification with spatial, temporal and textual features.

2. Literature Review

2.1. Social Media for Disaster Events

Social media data presents many advantages, such as timeliness of information, relevance at the community level, low cost, and adaptability (Keim and Noji 2011), over standard communication methods during disaster events (Houston et al. 2015). As a result, they are widely used for real-time dissemination of information by allowing for both sending and receiving of messages during disaster events (Xiao, Huang, and Wu 2015; Keim and Noji 2011). When traditional sources of communication lack information or cannot keep up to date on current information, social media can also serve as a backchannel communication platform allowing user-driven information acquisition and sharing (Xiao, Huang, and Wu 2015; Sutton, Palen, and Shklovski 2008). Peer-to-peer backchannel communications on social media fill information gaps when official sources of information are unavailable.

Meanwhile, social media is also widely leveraged for disaster event detection (Ford 2011), SA establishment (Huang and Xiao 2015), and disaster mapping (Li et al. 2018). 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 assistance on social media. For instance, during a 2011 tsunami off the coast of Japan, several tweets were direct requests for assistance (Acar and 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 and 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 SA for coordination and relief operations (Huang and 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. Zahra et al. (2017) investigated different types of sources on tweets related to eyewitnesses and classifies them into three types (i) direct eyewitnesses, (ii) indirect eyewitnesses, and (iii) vulnerable eyewitnesses.

While the use of social media for disasters clearly has a variety of advantages over traditional methods of communication, social media has raised concerns to the veracity of its data and grand challenges while being used to make decisions (Goodchild and Glennon 2010; Goodchild and 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 he or she experienced at a different time and location (Gao, Barbier, and Goolsby 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). 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. Another problem with the veracity of social media data is when social media is used maliciously (Huang and Xiao 2015; Yang et al. 2019). 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 easily (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. (2015) proposed a simple three-phase disaster classification for social media, pre-event, event, and post-event. During the pre-event phase, social media users send and receive information about the disaster event. 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 during different disaster phases (Huang and Xiao 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 generated in real-time and at changing rates. Among the massive data generated during a disaster event, only a small portion contributes to the establishment of SA. Any solution for utilizing relevant social media data during disaster events must have the processing capabilities to handle the stream of data efficiently. The proposed method will overcome the challenges of data volume by using text mining techniques to automatically search through the social media data for SA relevant information.

2.2. Text Mining for Extracting Disaster Relevant Information

One way to overcome the challenges associated with the volume of social media data is by searching through the text for patterns in the words that might signify data related to a disaster. Many different techniques have been applied for text mining social media data by developing 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. 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 and Liang 2012). 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, and Daelemans 2001). Hence, when creating a model for disaster events, unigrams are standard practice (Ashktorab et al. 2014; Spinsanti and Ostermann 2013; Huang and Xiao 2015).

Finally, a classification algorithm runs using training data. Traditionally, five machine learning algorithms, K-nearest neighbors, decision trees, naïve Bayes (Zahra, Ostermann, and Purves 2017), logistic regression (LR), SVM, and Random Forest (Zahra, Imran, and Ostermann 2020), are commonly used for text mining (Ashktorab et al. 2014; Spinsanti and Ostermann 2013; Huang and Xiao 2015; Bruns and Liang 2012; Zahra, Ostermann, and Purves 2017). However, recent studies indicate that deep learning based methods achieved better performance than these algorithms and in various natural language processing tasks (Yu et al. 2019). For example, a toponym recognition model, extending a general bidirectional recurrent neural network model, is developed for accurate location recognition in social media messages with various language irregularities (Wang, Hu, and Joseph 2020). 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, and also compared the performance of deep learning based CNN models, with several traditional classification models for SA information classification.

2.3. Remote Sensing for Tracking Disaster Events

Remote sensing data provides additional geographic information for detecting and tracking 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 and 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 and Smith 2009). Moreover, data mining remote sensing images 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 and Smith 2009). The proposed a context-aware text mining method 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, Twitter data estimated trajectories of earthquakes, tracked the locations of tornadoes, and detected wildfire hotspots (Crooks et al. 2013; Jain 2015; De Longueville, Smith, and Luraschi 2009). Meanwhile, disaster detection is not

always possible using social media as was determined during a 2013 flooding event (Fuchs et al. 2013).

One promising 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. 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. Similarly, Spinsanti and Ostermann (2013) introduced an approach that first geo-references and retrieves content from social media data, followed by an enrichment with additional geographic context information from authoritative data sources, and clustering spatio-temporally to support filtering and verification.

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 (Wang et al. 2018; Huang, Wang, and Li 2018a). For example, social media data was used to verify the presence of water in a specific area during a flooding event when remote sensing imagery was unavailable (Schnebele and Cervone 2013). Alternatively, Huang et al. (2018) introduced an approach to retrieve near real-time flood probability map by integrating the post-event remote sensing data with the real-time tweets (Huang, Wang, and Li 2018a). A flood inundation reconstruction model was further proposed to enhance the normalized difference water index derived from remote sensing imagery with both stream gauge readings and social media messages (Huang, Wang, and Li 2018b). Rosser et al. (2017) fused remote sensing, social media and topographic data sources for rapidly estimating flood inundation extent by a Bayesian statistical model to estimate the probability of flood inundation through weights-of-evidence analysis.

To sum up, much progress has been made to extract useful social media information, enrich them with additional authoritative datasets (de Albuquerque et al. 2015; Spinsanti and Ostermann 2013), and overcome the limitations of or enhance remote sensing data with social media data (Wang et al. 2018; Huang, Wang, and Li 2018a), (Rosser, Leibovici, and Jackson 2017; Schnebele and Cervone 2013). Given the infancy of spatial and social media data fusion for disaster management, no prior work fully examines the temporal, and particularly spatial features, and incorporates these features for extracting disaster relevant social media data. As such, this paper presents the context-aware text mining method, including the methodology for both social media and spatial data extraction using data mining algorithms.

3. Geographic Context-aware Text Mining

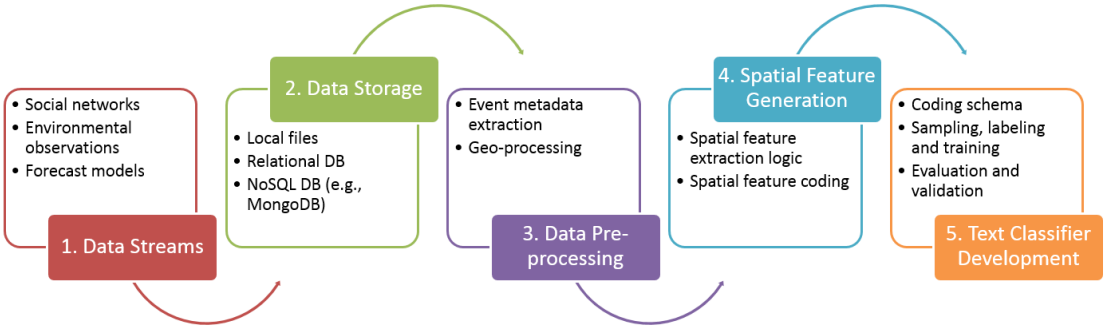

Figure 1. The workflow of geographic context-aware text mining

Context-aware text mining is an enhanced text mining method that incorporates SA about the disaster in the form of geographic data into the text classification. The proposed method includes five key components (Figure 1):

1. Social media and spatial data streams: A real-time disaster management scenario is the intended use of the method. Data in a variety of formats (e.g. point, polygon, raster, text) and volumes stream into the workflow from multiple sources.
2. Database storage: With the high volume of data generated dynamically, storage is necessary before and after processing. Due to scalability and spatial functionality, the storage of social media and spatial data is separate. NoSQL databases (e.g., MongoDB) can handle the high volume of social media content, while PostgreSQL/PostGIS supports a variety of spatial queries and operations of spatial data.
3. Data 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 method where event detection takes place. Disaster event detection is an important step for creating a dynamic spatial filter within the method compared to a traditional bounding box. In turn, the method 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 method 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 classifier development: Staying consistent with current methods, social media data with geographic metadata travel through a classifier to determine disaster relevant social media posts. This paper evaluates and validates the results from the classifier to determine if the addition of geographic information improves the text classification.

3.1. Data Streams, Processing, and Storage

3.1.1. Case study

To test the context-aware text mining method, Hurricane Sandy from 2012 was selected for the case study. Hurricane Sandy (October 22, 2012 – November 2, 2012) was the 18th named tropical cyclone for the 2012 Atlantic Hurricane. It made landfall in the United States (US) 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 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 et al. 2013). Rainfall is another impact from hurricanes that can lead to flooding, especially in urban and low-lying areas. The heaviest rain occurred in parts of Maryland, Virginia, and Delaware receiving between five and seven inches. The meteorological impact that resulted in the greatest causalities 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.

3.1.2. Data and Data Processing

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

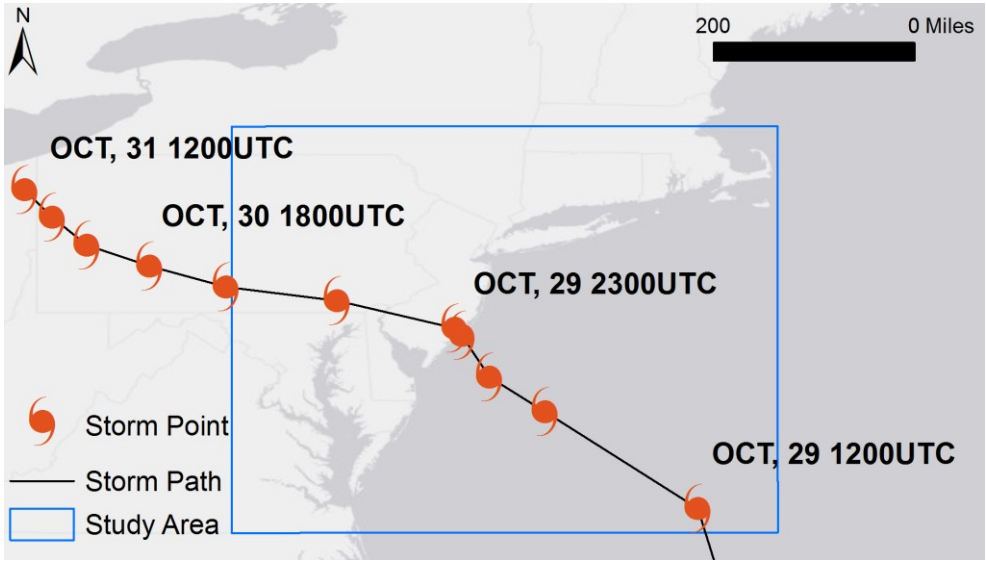

Figure 2. Study area and storm track of Hurricane Sandy

For the case study, Twitter, a text-based microblog with millions of active users, is selected as social media service for disaster relevant message extraction. 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 up to 1% of the most recent tweets based on search keywords and spatial extent (Morstatter et al. 2013). In total, 12.3 million geo-tagged Tweets were collected from October 28, 2012 to November 7, 2012 (Table 1) using Twitter streaming API with the "Hurricane" and "Sandy" as search keywords, and global as the geographical boundary. After performing a spatial filter based on the bounding box (Figure 2), the number of Tweets was reduced to 2.8 million.

The meteorological data offers a variety of ways to measure the hurricane both qualitatively and quantitatively. Additionally, the generated data is from different sources with different formats and different spatial and temporal resolutions (Table 1). Hurricane track points, produced by the National Hurricane Center, indicate the location of the hurricane center 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. To get more detailed weather measurements on the ground, 128 Automated Weather Observing System (AWOS) stations were used. These stations are primarily located at airports and take measurements at least every hour depending on the conditions. AWOS units collect data on many weather variables, such as wind, temperature, dew point, precipitation, and pressure. The radar data comes from six different NWS radar stations in the study area. In terms of data storage, the radar data was by far the largest dataset at ~45GB due to the temporal resolution (Table 1). Two products were kept for the rest of the study: base reflectivity (the common weather radar view) and storm total precipitation.

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 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 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 (Table 2) is one well-known way to accomplish this. Storm reports from the NWS were also 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 by in person visits, and updates storm reports frequently during a real disaster event.

The North American Mesoscale Forecast System (NAM) is a high-resolution forecast of hundreds of products. The 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 radar reflectivity. Finally, 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).

Table 2. Hurricane Sandy Watch/Warning/Advisory definitions (NWS 2021)

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

One challenge of using the context-aware text mining method is the data variety. 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 spatial data are transformed into the WGS84 coordinate reference system for data analysis. Second, vector data are stored in a Shapefile format and raster data in a GeoTIFF format. 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 dataset presented its own challenges for standardization. For example, the radar data exists natively in a binary format. The Weather and Climate Toolkit (Ansari, Del Greco, and Hankins 2010), developed by the National Oceanic and Atmospheric Administration (NOAA), read and exported the data as GeoTIFFs. Additionally, the radar and NAM model data were originally stored as single band GeoTIFFs for each product. GDAL (GDAL 2021) was used to merge 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 Python package METAR (Pollard 2021), each observation was decoded.

For the purposes of this study, the hurricane track points simulated the event detection phase of the context-aware text mining method (Table 1). With proper domain knowledge of the event detection algorithms, one could implement this step in the workflow.

3.1.3. Data Storage

The variety of social media data poses a challenge for traditional data management following the relational model (Huang and Xu 2014). Social media services utilize the NoSQL database model as a way to best manage their data. Unlike the traditional relational model, NoSQL implements many different data structures, such as document, graph, or key-value. The flexibility with NoSQL allows for data from multiple social media services to be stored in one location within the workflow. MongoDB (Banker 2011) was selected as the NoSQL database. In addition to the reasons state above, MongoDB stores its data in JavaScript Object Notation (JSON) which allows non-uniform fields to be added with no limitations. Most popular programming languages also easily parse JSON. Additionally, MongoDB is scalable allowing multiple servers to store and access the database. For the meteorological data, PostgreSQL was selected as PostgreSQL with the PostGIS extension can store both raster and vector data types, is open source, and provides a wide range of spatial functionality. Note other database systems could provide similar supports for social media data or spatial data management. For example, PostgreSQL also supports JSON data type and offers sufficient JSON operators and function to enable the storage of social media data.

3.2. Spatial Feature Generation

3.2.1. Spatial Feature Determination

The key step in the context-aware text mining method is the spatial feature generation. In this step, the spatial 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 and 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 and storm surge often have a dangerous impact. Another derived feature added was 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 added spatial feature. Two temporal features were added to reveal temporal detail to the feature space, including (1) the date of the social media post, and (2) a binary indication of whether the location was currently experiencing the storm or the storm had past.

3.2.2. 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. Given the spatiotemporal complexity of the datasets being used to generate features, ascribing context at the single time of a social media post requires more than a binary logic. 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 was to access the data from various sources. Since data was generated 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 exact spatial and temporal features were chosen based upon accuracy and temporal frequency limitations of the datasets selected as well as physical characteristics of the disaster event. For instance, the same logic used during a tornado event which occurs on a short time scale and smaller area would not provide the appropriate context to the social media post.

The last step in the feature generation algorithm was to rank the data products based on reliability of the data. Each spatial feature had multiple meteorological data sources 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

3.3. Annotation and Classification

3.3.1. 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, and Goolsby 2011; Huang and Xiao 2015; Imran et al. 2013; Vieweg et al. 2010). 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 further classified into more 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 SA. Based on Imran's (2013) coding schema, a majority of messages are classified as CA. Therefore, in our work, we created an additional class, named as Infrastructure and Resource (IR) containing messages reporting the status of infrastructure (e.g., transportation) and resource (e.g., gas, power, internet, food), which is not reported by an official news source or requests for donation or aid (Table 4).

Table 4. Social media classification scheme

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.
Infrastructure and Resource (IR)	Information related to IR that is not reported by an official news source or request for donation or aid	all metro service suspended until further notice yikes; NY subways scheduled close 7pm tonight; Gas station lines crazy
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. 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. Distances from the storm were clustered using the density-based spatial clustering of applications with noise (DBSCAN) 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 (Scale 2021b)
	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 Scale (NHC 2017). This scale is represented by the storm category feature. Another useful meteorological scale is the Beaufort Wind Scale (Scale 2021a). 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 Beaufort Wind Scale. Precipitation intensity feature is calculated using the NWS Radar Scale, 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.

3.3.2. 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, LR, and neural networks. In particular, the Naïve Bayes and support-vector machines models are commonly used (Ashktorab et al. 2014; Spinsanti and Ostermann 2013; Huang and Xiao 2015; Takahashi, Tandoc Jr, and Carmichael 2015). However, deep learning has produced better results for various tasks in text mining, such as topic classification, sentiment analysis, question answering, and language translation (Yu et al. 2019). As such, this study will examine the capability of CNN models, one of the popular deep learning based methods, for SA message classification.

Figure 3 presents the CNN architecture, which is the configuration for tweet message classification. Preprocessing converts tweets into lists of 50 integers and represents each word of the tweet by an integer. The preprocessed tweet then passes through the first layer, word embedding, which expands the word integers to a larger matrix and represents them in a more meaningful way. The word embedding layer uses Word2Vec (Mikolov et al. 2013) to embed semantic similarity information in the representation of words and expands each word into a vector of 300. The convolution layer extracts features from the word embedding and transforms them through global max pooling. The convolution layer uses neurons with filter size of 3, stride of 1, zero padding, and depth of 250 (Table 6). The extracted features are then concatenated with spatial and temporal features, which represent the tweet's distance from hurricane center and the surrounding geographic and meteorological environment. Then two fully connected layers predict the themes of each tweet. Dropout layers are utilized before the convolution layer and the last fully connected layer, while activation functions are used after the convolution layer and the fully connected layers.

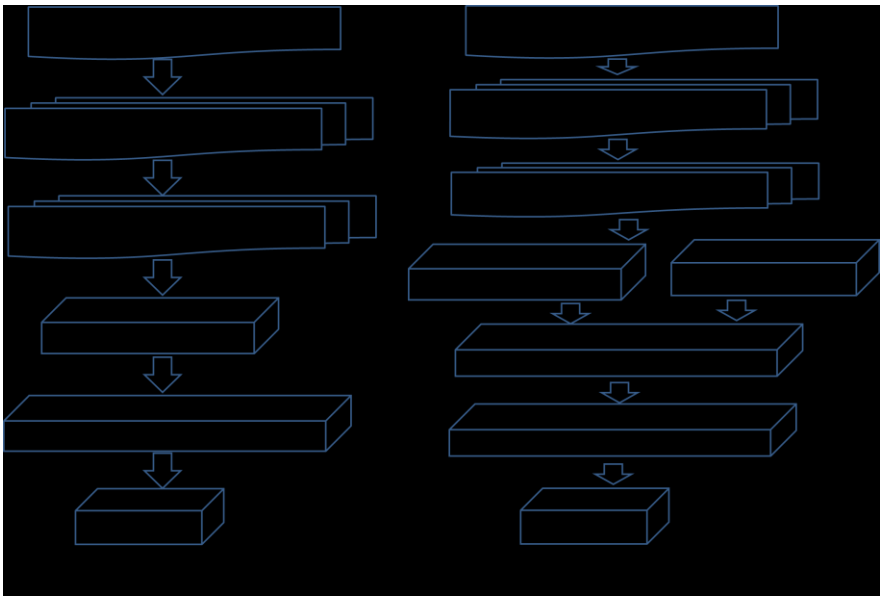

Figure 3. Overall architecture of the CNN: a) tweet text only, and b) using spatially enhanced tweets.

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

Table 6. Dimensions of layers and operations

Layer	Operator	Output Height	Output Width	Output Depth
Input	-	1	50	-
Embedding		1	50	300
Dropout	Rate = 0.25	1	50	300
Convolution	Stride=1, Zero padding =0, Depth = 250, Filter size = 3; Activation = ReLU	1	50	250
Global max pooling	-	1	1	250
Metadata*	-	1	1	7
Concatenate*	-	1	1	257
Fully connected	Output depth = 250; Activation = ReLU	1	1	250
Dropout	Rate = 0.25	1	1	250
Fully connected	Output depth = 5; Activation = Softmax	1	1	5

**Metadata and concatenate layers only exist for environmental enhanced tweet classification architecture (Figure 3b). These layers are eliminated for tweet text only classification architecture (Figure 3a).*

One limitation of the concatenation between the spatiotemporal features and the output of the global max pooling layer is that the dimension of the max pooling layer is significantly higher than the one of the spatiotemporal features. An alternative approach is to assign the spatiotemporal features with a higher degree of influence using higher weights.

3.4. Spatiotemporal Distribution of Sample Data

Before running the classification experiment (Section 4), understanding the spatiotemporal distribution of the labelled sample data was important for interpreting the results. Of the 5,000 social media posts sampled, 1,920 posts were labelled as informative. Within the sample, only 2 posts belong to the People missing or found class, and therefore this class was removed from the classification experiments due to the low sample count.

Shown as Figure 4, the breakdown in posts per class were not surprising. CA was the most general class and contained the most posts, followed by the IR class. The CA posts include disaster preparation information (e.g., stock up food) and status (e.g., wind speed of the Hurricane Sandy) of a disaster event. During the disaster, the IR 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 samples without a clear temporal pattern. Analysis of the IS class is included in Section 4.

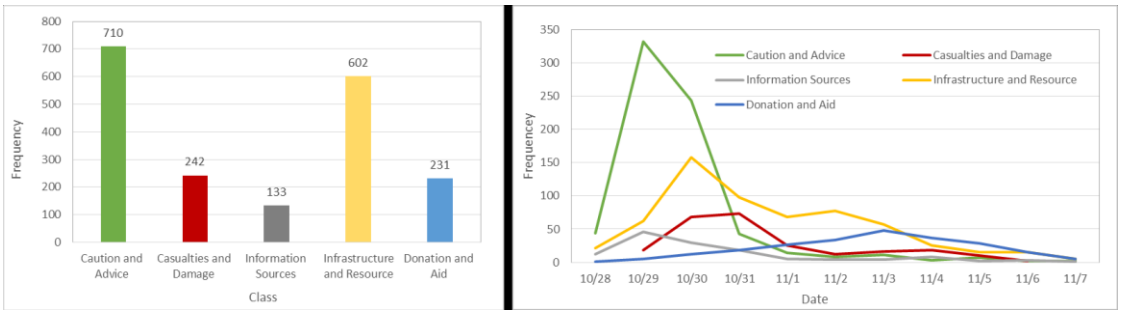

Figure 4. The frequency (Left) and temporal distribution (Right) of informative posts

4. Experimental Design and Analysis

This section first describes the experimental design to identify spatial and temporal features for the Hurricane Sandy case study that are key to describing and then classifying disaster relevant social media posts, followed by the results from the spatial and temporal feature determination. Finally, the section discusses the results of the feature combination experiments.

4.1. Experimental Design

4.1.1. Selected Experiments

To determine if geographic features improve the accuracy of current text-only methods, a series of experiments were designed (Table *7*). A text-only experiment would serve as the control to the other spatial experiments. The next experiment used all features. However, utilizing all features poses two problems. First, using too many features, especially irrelevant and correlated ones, may not necessarily generate desired model (Kohavi and Sommerfield 1995). Second, the goal of this research is to identify the key spatial features or the minimal number of features that give the best accuracy. As such, four additional experiments tested the number of features based on hypotheses derived from domain knowledge.

Table 7. Context-aware text mining experiment design

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

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 5). Since the machine learning algorithms selected for this study are dependent on probabilities, the expectation is that the date feature will improve the classification. In addition, combining the distance and date feature with the text feature would likely give the best results. Out of curiosity, a third experiment related to distance and time was created which combined what the authors 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. Running the labelled data through a supervised classifier involves taking a subset of the data to train the classifier and the remainder of the labelled data to test the classifier. One potential error associated with this approach is for bias to exist in the training data resulting 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. Meanwhile, 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 using only the text feature to select the most effective classification algorithm, which was then used for the remainder of experiments. As discussed earlier, from the well-established classification algorithms, three were selected based on previous text mining studies: LR, SVM, and CNN.

Not surprisingly, the CNN performed better in text mining tasks demonstrated with higher accuracy values than the other two algorithms in all three evaluation metrics (Table 8). Therefore, the remaining experiments utilized the CNN.

Table 8. Text-only algorithm experiment

Algorithm	Precision	Recall	F1-score
CNN	**0.76**	**0.74**	**0.76**
LR	0.56	0.53	0.50
SVM	0.70	0.69	0.69

4.2. Feature Determination

4.2.1. Spatial Features

Determining the feature annotation for the distance from the center of the disaster is not a trivial task (Section 3.5.1). 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, the distance was coded 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. In fact, the sample data mostly 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 the DBSCAN algorithm (Ester et al. 1996) to cluster the posts based on distance from the center (*l*), the date (*d*), and the latitude (*lat*) and longitude (*lon*). DBSCAN is a density-clustering algorithm which groups points based on a minimum number of points (i.e., samples; *minpts*) within a specified epsilon or search distance (*eps*). High-density points cluster into groups and low-density points are treated as noise. After running DBSCAN, a set of clusters are produced with each cluster and we can assign each cluster with a unique ID. Next, each clustered sample point is given its associated cluster ID as the distance feature, and the distance feature for all noisy points is set to 0. To test which attributes (*l*, *d*, *lat* and *lon*), should be used to calculate distanced between samples in DBSCAN, we evaluate the classifiers generated by using the clustering results with different combinations of these attributes, including 1) *l*, 2) *l* and *d*, 3) *lat* and *lon*, and *d,* and 4) *lat* and *lon*, along with text as the input features. It is worth noting that the results of DBSCAN are sensitive to the choice of *minpts* and *eps* value. In our work, *minpts* value remained at 4 and the *eps* value was changed until a consistent number of clusters were found for each experiment setting. The highest F-score occurs when *l* and *d* are the clustering attributes. However, using the date feature in the DBSCAN and as a feature in the classification (Table 8) could result in feature redundancy. To minimize feature redundancy, the second-best trial, *lat*, *lon*, and *l*, was selected as the classification of the distance feature.

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. For example, a day before the disaster event, the few messages posted primarily related to the CA and IR class. This type of result meets the expectation of a disaster because no damage exists prior to the event.

Figure 5. 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 5a). As the storm affects the disaster area, the IS and CD classes become more prominent (Figure 5b). 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 5c). The posts after the disaster occur in areas significantly impacted by the storm.

4.3. Feature Combination Performance

To determine which feature or features best describe the disaster in the classification models and examine whether the addition of features improves the accuracy of the classification, different feature combination experiments are performed based on the CNN models. The 10-fold cross-validation results indicate that each experiment, which combined spatial and temporal feature with text, improves the overall accuracy compared to the text-only classification by 4 percentage points on average (Table 9). The enhanced proximity and time experiment has the highest average accuracy (0.81) and F-score (0.80) of the experiments. Adding meteorological features and adding distance and time, have a similar amount of improvement to the recall and F-score (Table 9). The results from these experiments indicate proximity to the disaster is the best feature to describe the disaster event.

Table 9. Feature combination experiment results

Experiment	Class	Precision	Recall	F-score
Text-only	CA	0.65	0.87	0.75
	IR	0.66	0.64	0.65
	CD	0.58	0.38	0.46
	IS	0.93	0.72	0.81
	DA	0.89	0.67	0.76
	Averages	**0.76**	**0.74**	**0.74**
All	CA	0.81	0.77	0.79
	IR	0.63	0.85	0.72
	CD	0.76	0.53	0.62
	IS	0.85	0.85	0.85
	DA	0.86	0.86	0.86
	Averages	**0.80**	**0.80**	**0.80**
Proximity	CA	0.81	0.78	0.79
(text, distance)	IR	0.63	0.72	0.67
	CD	0.79	0.63	0.70
	IS	0.86	0.87	0.86
	DA	0.80	0.84	0.82
	Averages	**0.80**	**0.80**	**0.80**
Proximity & Time	CA	0.75	0.82	0.78
(text, distance, date)	IR	0.65	0.70	0.67
	CD	0.77	0.50	0.61
	IS	0.85	0.85	0.85
	DA	0.86	0.76	0.81
	Averages	**0.79**	**0.78**	**0.78**
Enhanced Proximity & Time	CA	0.82	0.75	0.79

(text, distance, date, disaster status)	IR	0.62	0.92	0.74
	CD	0.78	0.58	0.67
	IS	0.86	0.82	0.84
	DA	0.86	0.86	0.86
	Averages	**0.81**	**0.80**	**0.80**
Meteorological	CA	0.73	0.84	0.78
(text, storm category, precipitation, wind, flood, warning)	IR	0.72	0.54	0.61
	CD	0.63	0.61	0.62
	IS	0.87	0.85	0.86
	DA	0.90	0.79	0.84
	Averages	**0.79**	**0.79**	**0.78**

Assessing individual classes reveals a few patterns that exist in the different experiments. First, adding spatial and temporal features had a large impact on the CA (caution and advice) with an increasing of accuracy and decreasing recall, indicating that more messages (both true and false positives) are classified as CA. Second, the precision of IR category stayed relatively consistent for each experiment. However, the recall increased with the addition of distance and time features, and yet dropped after adding the meteorological features. Third, in general the DA (damage and donation) classes improved dramatically in the recall resulting in much higher F-scores. For disaster relevant information retrieval, identifying a greater number of disaster relevant posts accurately or increasing the recall is important. In particular, including a distance feature all improves recall significantly, which makes sense given the class definitions and sensitivity to space. Finally, the significant decreasing of precision and increasing of recall indicate that adding addition features can largely reduce false positives for the IS class.

5. Conclusion and Future Work

This paper presents an enhanced text mining method that considers geographic context by incorporating spatial data into the classification of disaster-relevant social media posts. To extract SA information, current approaches focus on matching or text mining keywords to classify disaster relevant posts. Given the high degree of variability in natural language processing, current methods are not without error and have difficulty understanding context. Specifically, the proposed method ingests and processes spatial data that can inform the geographic context of the disaster at a given place in time to enhance the text data by providing additional SA. Then the method combines the spatial data with the text data and classifies the posts based on relevance using a CNN. Data from Hurricane Sandy provided a means for testing the method. The disaster presents several big data challenges in data volume and variety. Collection, processing, and standardization varied for each spatial data set. Using fuzzy logic, spatial features (e.g. distance, wind, flood, and precipitation) were bound to each social media post in the sample data set.

A group of experiments were designed to test the performance of the spatial and temporal features with the text feature. Based on the results, the addition of spatial and temporal features to the text classification did improve the overall classification accuracy compared to current methods. Moreover, 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. From the experiments, the impact of spatial features on individual classes was relative to the class definition. The additional features had significant impact on classes (e.g. casualties and damage or donation and aid) whose class definitions involved spatial and temporal components.

Currently, research of spatial feature creation and definition is non-existent. 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. In fact, a wide range of diversity of spatial data and varying approaches can be leveraged to derive and code spatial and temporal features, represented as coded numeric values in the classification models , and some of the data are highly domain (or disaster type) specific. This paper defined and characterized several key types of spatial features (e.g., rain, wind, and pressure) for the task of social media message classification in hurricanes characterized by heavy rain, high winds, and low atmospheric pressure. Meanwhile, during an earthquake event, seismic activities (e.g., seismic waves, geographic coordinates of its epicenter, depth of the epicenter) in the area are more useful. As such, more investigation into optimal spatial feature creation from multi-sourced spatial data for different types of disaster events is a possible step in future to improving the classification results given the diversity of spatial data and approaches to derive and code features.

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 combines the predictions of several estimators for a given algorithm. Since quick response is important in disaster management, a framework based on unsupervised learning (Zhou et al. 2021), and self-learning methods (Peng et al. 2020; Peng, Huang, and Rao 2021), avoiding manual labelling process would be more ideal for our task. Additionally, 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. In addition to spatial features, visual features extracted from images associated with the social media posts, could also be further integrated to enhance the extraction of disaster relevant messages (Huang et al. 2019). Further, our labelled datasets include mostly CA (Caution and Advice) and IR (Infrastructure and Resource) categories, leading to class-imbalance (i.e., some classes have much fewer training samples than the other classes). In fact, class-imbalance has been widely acknowledged as one of the most challenging problems in machine learning and therefore well addressed in the literature (Sahare and Gupta 2012). A potential future work could employ different methods (e.g., weighing the classes differently) to improve class imbalance.

Finally, social media is not going away anytime soon. For disaster applications related to social media data, this study indicates the need for incorporating spatial data sources to improve information retrieval and provide greater SA. Existing studies indicated a high performance for a classification model based on traditional machine learning algorithms (e.g., Naïve Bayes) even when applied to data from a different geographic region with significant differences in social media user and usage characteristics (Zahra, Ostermann, and Purves 2017). Further, deep learning based models pre-training using Twitter data from past events have been demonstrated with higher performance while classifying tweet topic for an upcoming event to establish SA compared with these traditional classification models (Yu et al. 2019). In other words, once we build a model with datasets from historical events, it can be applied to classify disaster relevant social media messages generated from a later event. Since spatial datasets for extracting spatial features are generated from real-time forecasting models (e.g., NAM) and physical sensing networks (e.g., weather station observations, remote sensing), this framework is applicable to support both real-time decision making and post disaster analysis as long as a workflow of automatic retrieval of relevant datasets is created. Finally, applications of big geosocial media data are increasing common throughout a range of activities beyond just disaster response, from urban planning to market research to political activism (Shelton et al. 2014).

Acknowledgements

The authors would like to thank the funding support from the Vilas Associates Competition Award at University of Wisconsin-Madison (UW-Madison), and the National Science Foundation under Grant 1940091.

Availability of data and materials

The datasets used and/or analysed during the current study will be available from online repositories and the corresponding author on request.

Competing interests

The authors declare that they have no competing interests.

References

Acar, Adam, and Yuya Muraki. 2011. "Twitter for crisis communication: lessons learned from Japan's tsunami disaster." *International Journal of Web Based Communities* 7 (3):392-402.

Alberts, Timothy A, Phillip B Chilson, BL Cheong, and RD Palmer. 2011. "Evaluation of weather radar with pulse compression: Performance of a fuzzy logic tornado detection algorithm." *Journal of Atmospheric and Oceanic Technology* 28 (3):390-400.

Ansari, Steve, S Del Greco, and B Hankins. 2010. The weather and climate toolkit. Paper presented at the AGU Fall Meeting Abstracts, San Francisco, California, 13–17 December 2010.

Ashktorab, Zahra, Christopher Brown, Manojit Nandi, and Aron Culotta. 2014. Tweedr: Mining twitter to inform disaster response. Paper presented at the 11th International Conference on Information

Systems for Crisis Response and Management (ISCRAM 2014), Pennsylvania, USA.

Bakillah, Mohamed, Ren-Yu Li, and Steve HL Liang. 2015. "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* 29 (2):258-79.

Banker, Kyle. 2011. *MongoDB in action*. Shelter Island, NY: Manning Publications Co.
Blake, ES, TB Kimberlain, RJ Berg, JP Cangialosi, and JL Beven II. 2013. "Tropical cyclone report: Hurricane Sandy." *National Hurricane Center* 12:1-10.
Bruns, Axel, and Yuxian Eugene Liang. 2012. "Tools and methods for capturing Twitter data during natural disasters." *First Monday* 17 (4).
Crooks, Andrew, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. 2013. "# Earthquake: Twitter as a distributed sensor system." *Transactions in GIS* 17 (1):124-47.
de Albuquerque, João Porto, Benjamin Herfort, Alexander Brenning, and Alexander Zipf. 2015. "A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management." *International Journal of Geographical Information Science* 29 (4):667-89.
De Longueville, Bertrand, Robin S Smith, and Gianluca Luraschi. 2009. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. Paper presented at the Proceedings of the 2009 international workshop on location based social networks, Seattle, Washington, 3 November 2009.
Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Paper presented at the the Second International Conference on Knowledge Discovery and Data Mining (KDD), Portland, Oregon, August 2-4.
Ford, R. 2011. "Earthquake: Twitter users learned of tremors seconds before feeling them." *The Hollywood Reporter* 23.
Fuchs, Georg, Natalia Andrienko, Gennady Andrienko, Sebastian Bothe, and Hendrik Stange. 2013. Tracing the German centennial flood in the stream of tweets: first lessons learned. Paper presented at the Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information, Orlando, Florida, 5 November 2013.
Gao, Huiji, Geoffrey Barbier, and Rebecca Goolsby. 2011. "Harnessing the crowdsourcing power of social media for disaster relief." *IEEE Intelligent Systems* (3):10-4.
GDAL. "GDAL Documentation." https://gdal.org/.
Goodchild, Michael F, and J Alan Glennon. 2010. "Crowdsourcing geographic information for disaster response: a research frontier." *International Journal of Digital Earth* 3 (3):231-41.
Goodchild, Michael F, and Linna Li. 2012. "Assuring the quality of volunteered geographic information." *Spatial statistics* 1:110-20.
Halteren, Hans van, Jakub Zavrel, and Walter Daelemans. 2001. "Improving accuracy in word class tagging through the combination of machine learning systems." *Computational linguistics* 27 (2):199-229.
Houston, J Brian, Joshua Hawthorne, Mildred F Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R Halliwell, Sarah E Turner McGowen, Rachel Davis, Shivani Vaid, and Jonathan A McElderry. 2015. "Social media and disasters: a functional framework for social media use in disaster planning, response, and research." *Disasters* 39 (1):1-22.
Huang, Qunying, Guido Cervone, Duangyang Jing, and Chaoyi Chang. 2015. DisasterMapper: A CyberGIS framework for disaster management using social media data. Paper presented at the Proceedings of the 4th International ACM SIGSPATIAL Workshop on Analytics for Big Geospatial Data, Bellevue, WA, USA, November 3 - 6, 2015.
Huang, Qunying, and Yu Xiao. 2015. "Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery." *International Journal of Geo-Information* 4 (3):19. doi: 10.3390/ijgi4031549.
Huang, Qunying, and Chen Xu. 2014. "A data-driven framework for archiving and exploring social media data." *Annals of GIS* 20 (4):265-77.

Huang, Xiao, Cuizhen Wang, and Zhenlong Li. 2018a. "A near real-time flood-mapping approach by integrating social media and post-event satellite imagery." *Annals of GIS* 24 (2):113-23.
———. 2018b. "Reconstructing flood inundation probability by enhancing near real-time imagery with real-time gauges and tweets." *IEEE transactions on Geoscience and Remote Sensing* 56 (8):4691-701.
Huang, Xiao, Cuizhen Wang, Zhenlong Li, and Huan Ning. 2019. "A visual–textual fused approach to automated tagging of flood-related tweets during a flood event." *International Journal of Digital Earth* 12 (11):1248-64.
Imran, Muhammad, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Practical extraction of disaster-relevant information from social media. Paper presented at the Proceedings of the 22nd international conference on World Wide Web companion, Rio de Janeiro, Brazil May 13 - 17, 2013.
Jain, Saloni. 2015. "Real-Time Social Network Data Mining for Predicting the Path for a Disaster." Georgia State University
Joachims, Thorsten. 1998. "Text categorization with support vector machines: Learning with many relevant features." *Machine learning: ECML-98*:137-42.
Keim, Mark E, and Eric Noji. 2011. "Emergent use of social media: a new age of opportunity for disaster resilience." *American journal of disaster medicine* 6 (1):47-54.
Kohavi, Ron, and Dan Sommerfield. 1995. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. Paper presented at the Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21.
Lakshmanan, Valliappa, and Travis Smith. 2009. "Data mining storm attributes from spatial grids." *Journal of Atmospheric and Oceanic Technology* 26 (11):2353-65.
Landwehr, Peter M, and Kathleen M Carley. 2014. "Social media in disaster relief." In *Data mining and knowledge discovery for big data*, edited by Wesley W. Chu, 225-57. Berlin, Heidelberg: Springer.
Li, Zhenlong, Cuizhen Wang, Christopher T Emrich, and Diansheng Guo. 2018. "A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods." *Cartography and Geographic Information Science* 45 (2):97-110.
Lindsay, Bruce R. 2011. "Social media and disasters: Current uses, future options, and policy considerations." In, 1-13. Washington, DC: Congressional Research Service, Report for Congress.
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Paper presented at the Advances in neural information processing systems.
Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. Paper presented at the the 7th International AAAI Conference on Weblogs and Social Media 28 Jun.
NHC. "National Hurricane Center." http://www.nhc.noaa.gov/.
NWS. "Definitions of Weather Watch, Warnings and Advisories." https://www.weather.gov/lwx/WarningsDefined.
Peng, Bo, Qunying Huang, and Jinmeng Rao. 2021. Spatiotemporal Contrastive Representation Learning for Building Damage Classification. Paper presented at the 2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Brussels, Belgium, July 11 – 16, 2021.
Peng, Bo, Qunying Huang, Jamp Vongkusolkit, Song Gao, Daniel B Wright, Zheng N Fang, and Yi Qiang. 2020. "Urban Flood Mapping With Bitemporal Multispectral Imagery Via a Self-Supervised Learning Framework." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14:2001-16.
Pollard, Tom. "METAR." https://pypi.python.org/pypi/metar.

Rosser, Julian F, DG Leibovici, and MJ Jackson. 2017. "Rapid flood inundation mapping using social media, remote sensing and topographic data." *Natural Hazards* 87 (1):103-20.
Roy, Chandan, and Rita Kovordányi. 2012. "Tropical cyclone track forecasting techniques—A review." *Atmospheric research* 104:40-69.
Sahare, Mahendra, and Hitesh Gupta. 2012. "A review of multi-class classification for imbalanced data." *International Journal of Advanced Computer Research* 2 (3):160.
Scale, Beaufort Wind. "Beaufort Wind Scale." https://www.weather.gov/tbw/beaufort.
Scale, NWS Radar. "The Front." https://www.weather.gov/media/publications/front/06nov_Front.pdf.
Schnebele, Emily , and Guido Cervone. 2013. "Improving remote sensing flood assessment using volunteered geographical data." *Natural Hazards and Earth System Sciences* 13 (3):669-77.
Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook. 2014. "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'." *Geoforum* 52:167-79.
Spinsanti, Laura, and Frank Ostermann. 2013. "Automated geographic context analysis for volunteered information." *Applied Geography* 43:36-44.
Sutton, Jeannette, Leysia Palen, and Irina Shklovski. 2008. Backchannels on the front lines: Emergent uses of social media in the 2007 southern California wildfires. Paper presented at the Proceedings of the 5th International ISCRAM Conference.
Takahashi, Bruno, Edson C Tandoc Jr, and Christine Carmichael. 2015. "Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines." *Computers in Human Behavior* 50:392-8.
Tobler, Waldo R. 1970. "A computer movie simulating urban growth in the Detroit region." *Economic Geography* 46 (sup1):234-40.
Vieweg, Sarah, Amanda L Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems, Montréal Québec, Canada, April 22 - 27, 2006.
Wang, Han, Erik Skau, Hamid Krim, and Guido Cervone. 2018. "Fusing heterogeneous data: A case for remote sensing and social media." *IEEE transactions on Geoscience and Remote Sensing* 56 (12):6956-68.
Wang, Jimin, Yingjie Hu, and Kenneth Joseph. 2020. "NeuroTPR: A neuro‐net toponym recognition model for extracting locations from social media messages." *Transactions in GIS* 24 (3):719-35.
Xiao, Yu, Qunying Huang, and Kai Wu. 2015. "Understanding social media data for disaster management." *Natural Hazards* 79 (3):1663-79. doi: 10.1007/s11069-015-1918-0.
Yang, Jingchao, Manzhu Yu, Han Qin, Mingyue Lu, and Chaowei Yang. 2019. "A twitter data credibility framework—Hurricane harvey as a use case." *ISPRS International Journal of Geo-Information* 8 (3):111.
Yu, Manzhu, Qunying Huang, Han Qin, Chris Scheele, and Chaowei Yang. 2019. "Deep learning for real-time social media text classification for situation awareness–using Hurricanes Sandy, Harvey, and Irma as case studies." *International Journal of Digital Earth* 12 (11):1230-47.
Zahra, Kiran, Muhammad Imran, and Frank O Ostermann. 2020. "Automatic identification of eyewitness messages on twitter during disasters." *Information processing & management* 57 (1):102107.
Zahra, Kiran, Frank O Ostermann, and Ross S Purves. 2017. "Geographic variability of Twitter usage characteristics during disaster events." *Geo-spatial Information Science* 20 (3):231-40.
Zhou, Sulong, Pengyu Kan, Qunying Huang, and Janet Silbernagel. 2021. "A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura." *Journal of Information Science*:01655515211007724.

Zou, Lei, Nina SN Lam, Heng Cai, and Yi Qiang. 2018. "Mining Twitter data for improved understanding of disaster resilience." *Annals of the American Association of Geographers* 108 (5):1422-41.