Development of a National-Scale Big Data Analytics Pipeline to Study the Potential Impacts of Flooding on Critical Infrastructures and Communities

Nattapon Donratanapat

Follow this and additional works at: https://scholarcommons.sc.edu/etd

Part of the Computer Engineering Commons

Recommended Citation

This Open Access Thesis is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact dillarda@mailbox.sc.edu.
Development of a national-scale big data analytics pipeline to study the potential impacts of flooding on critical infrastructures and communities

By

Nattapon Donratanapat

Bachelor of Science
University of South Carolina, 2018

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in

Computer Engineering

College of Engineering and Computing

University of South Carolina

2019

Accepted by:

Jose Vidal, Director of Thesis
Qiang Zeng, Reader
Seeyedehzahra "Vidya" Samadi, Reader
Cheryl L. Addy, Vice Provost and Dean of the Graduate School
DEDICATION

This thesis is proudly dedicated to my family: Suphakarn Khammuangjai, Darunee, and Nopadon Donratanapat, and Brandon, Rosy, and Logan Granger, and for their support, pray, love, and care.
ACKNOWLEDGEMENTS

This project represents the result of hard work and dedication contributed by many different people.

My Mentor, Dr. Vidya Samadi and Dr. Jose Vidal, for conscientious guidance and encouragement to accomplish this assignment along with their valuable advice.

Jeremy Lane, who gave his knowledge of data science throughout the length of this project, and his support for completion of this project.

My parents, who provided me with valuable advice in life

The most deserving of my gratitude is my wife, Prim, who helps and supports me through this project. I would not have been able to overcome the tough times such a venture entails without her solidarity and support.

Nattapon Donratanapat
07/31/2019, Columbia, South Carolina
ABSTRACT

With the rapid development of the Internet of Things (IoT) and Big data infrastructure, crowdsourcing techniques have emerged to facilitate data processing and problem solving particularly for flood emergencies purposes. A Flood Analytics Information System (FAIS) has been developed as a Python Web application to gather Big data from multiple servers and analyze flooding impacts during historical and real-time events. The application is smartly designed to integrate crowd intelligence, machine learning (ML), and natural language processing of tweets to provide flood warning with the aim to improve situational awareness for flood risk management and decision making. FAIS allows the user to submit search request from the United States Geological Survey (USGS) as well as Twitter through a series of queries, which is used to modify request URL sent to data sources. This national scale prototype combines flood peak rates and river level information with geotagged tweets to identify a dynamic set of at-risk locations to flooding. The list of prioritized areas can be updated every 15 minutes as the crowdsourced data and environmental information and condition change. In addition, FAIS uses Google Vision API (application programming interface) and image processing algorithms to detect objects (flood, road, vehicle, river, etc.) in time-lapse digital images and build valuable metadata into image catalog. The application performs Flood Frequency Analysis (FFA) and computes design flow values corresponding to specific return periods that can help engineers in designing safe structures and in protection against economic losses due to
maintenance of civil infrastructure. FAIS is successfully tested in real-time during Hurricane Dorian flooding event across the Carolinas where the storm made extensive damage and disruption to critical infrastructure and the environment. The prototype is also verified during historical events such as Hurricanes Matthew and Florence flooding for the Lower PeeDee Basin in the Carolinas.
# TABLE OF CONTENTS

DEDICATION ........................................................................................................ iii

ACKNOWLEDGEMENTS ................................................................................... iv

ABSTRACT ........................................................................................................... v

LIST OF FIGURES ............................................................................................... ix

SOFTWARE AND DATA AVAILABILITY ......................................................... x

CHAPTER 1 INTRODUCTION AND BACKGROUND .................................... 1

1.1 Motivation and Significance ........................................................................ 2

1.2 Research Questions ...................................................................................... 2

CHAPTER 2 LITERATURE REVIEW AND RELATED RESEARCH WORKS ...... 5

CHAPTER 3 METHODOLOGIES ...................................................................... 8

3.1 Data Gathering and Analytics .................................................................... 8

3.2 Data Analytics Approaches ....................................................................... 9

3.3 Machine Learning and Image Processing Approaches .............................. 10

3.4 Flood Frequency Analysis (FFA) ................................................................. 12

3.5 Crowd-sourced Data Collection ................................................................. 13

3.6 Twitter APIs Tools ..................................................................................... 15

3.7 USGS API for Data Gathering and Web Cameras ...................................... 19

3.8 Web Development Framework .................................................................... 21

CHAPTER 4 DISCUSSION .................................................................................. 23
4.1 Applying and Testing FAIS during Real-time and Historical Flooding Events ........................................................................................................23

4.2 Description of Pipeline Steps and Design Options ........................................23

4.3 Applying and Testing FAIS Pipeline in Real-time During Hurricane Dorian 26

4.4 Applying and testing FAIS pipeline during historical events ................................27

4.5 Image Analysis Using Google Vision API and Image Processing Algorithm 28

4.6 USGS Data Collection ..........................................................................................30

4.7 Flood Frequency Analysis (FFA) .........................................................................31

4.8 Limitations of Our Prototype and Further Work ............................................33

CHAPTER 5 CONCLUSION .........................................................................................34

REFERENCES ........................................................................................................35
LIST OF FIGURES

Figure 3.1. Tweet Streaming Relationship .............................................................. 17

Figure 3.2 Universal Resource Locator for sending request to Twitter in JSON format . 18

Figure 3.3 Request URL for USGS Historical Data .................................................. 19

Figure 3.4 Request URL for USGS Real-time Data .................................................... 20

Figure 3.5 FAIS workflow and structure ................................................................. 22

Figure 4.1 Geotagged tweets retrieved via Twitter Search API in real-time during September 04, 2019 Hurricane Dorian event using specific location searches combined with keywords................................................................. 27

Figure 4.2 Geotagged tweets retrieved via Twitter Search API during Hurricane Florence event. .......................................................................................................................................................... 28

Figure 4.3 Flooded (right) and Not flooded (left) images with detected labels and scores during Hurricane Florence event at New Bern, NC. .................................................. 29

Figure 4.4 Flooded (right) and not flooded (left) bridges with detected labels and scores during Hurricane Florence event at New Bern, NC. .................................................. 29

Figure 4.5 Historical Data of flood station 04001000, in Michigan, from 08/01/2019 to 10/31/2019 .................................................................................................................................................. 30

Figure 4.6 North Carolina realtime flood data as of the October 18, 2019. .................. 31

Figure 4.7 South Carolina Realtime flood data as of the October 18, 2019 ............... 31

Figure 4.8 Flood frequency analysis for USGS 02172035 .............................................. 32
SOFTWARE AND DATA AVAILABILITY

Flood Analytic Integration System (FAIS): Various Scripts as they apply to pipeline development

Description: Data Gathering application designed for analyzed flood and Twitter data.

Developer: Nattapon Donratanapat

Contact: nattapon@email.sc.edu, pleuk5667@gmail.com

Software Access: https://pypi.org/manage/project/fais/releases/

Year First Available: 2019

Hardware required: Windows, Linux, MacOS

Hardware required: Intel i3 or mid-performance PC with multicore processor and SSD main drive, 4Gb memory recommended.

Software availability: All source codes can be found at GitHub public repositories https://github.com/VidyaSamadi/Flood-Data-Analytics-App

Cost: Free. Software and source code are released under the Massachusetts Institute of Technology License.
CHAPTER 1

INTRODUCTION AND BACKGROUND

Python is an interpreted, high-level, general-purpose programming language which its popularity increases in science and education field due to its ease of use and implementation. It has a large community support with a wide amount and range of APIs that are publicly available for all developers to access. This allows python to be used in both the academia and industrial program development activities.

Flood Analytics information System as a web Python platform aims to help improve flood data gathering method for the researchers and potentially other users like scientists, news reporters, and data analytic personals. To analyze flooding impacts, a large volume of Big data must be collected from various servers and applications. FAIS allows the user to directly download flood related data from the United States Geological Survey (USGS) in the modified csv formatted and data frames. With the increase in importance of social media, our developed pipeline allows the user to query tweets from Twitter by a specific user and/or keyword. The tweet result can be download in a csv formatted file.

This dissertation is organized as follows. In section 2, the motivation of this research work is explained. The procedures and the functionality of the FAIS application are introduced and discussed in Section 3. Section 4 discusses the application and results of our product. Conclusions and future works are provided in Section 5.
1.1 **Motivation and Significance**

Floods are the most common and among the deadliest natural disasters in the United States. They cause damage and disruption to critical infrastructure, building, human, and the environment. To battle with this type of extreme event, many studies have been conducted to better understand flooding impacts and to identify the areas at-risk of flooding in real time from the latest national and environmental data sources. However, the exceptionally fast growth in the size and variety of flood-related datasets, especially real-time observations, poses a significant challenge to modeler and decision makers. Vast amounts of these data are collected on sub daily (e.g., 15 minutes) and daily bases by a wide range of sensors, cameras, and instruments (e.g., Demir and Krajewski, 2013). With the rapid development of earth observation technology and ground-based monitoring systems that produce time-lapse imagery and data, high spatial and temporal data have been recently applied for flood studies. Although, these Big data and images require innovative enabling technologies to improve the integration, retrieval, analysis and presentation of large amounts of information (Grolinger et al., 2013; Nativi et al., 2015) for flood disaster and risk analysis.

1.2 **Research Questions**

In this study, we developed and tested a prototype that combines machine learning, API gathering (both REST and Stream APIs), and geodata. FAIS is based on historical and real-time flood warnings and river level information, and natural language processing of tweets during flooding events, aiming to improve situational awareness for flood risk managers and other stakeholders. This involved novel automated data collection and
analysis of large volumes of tweets, geotagged data, and relevant social media feed. Developing FAIS required addressing three research questions which are:

1. How to programmatically identify at a national-scale (across the United States) areas at-risk of flooding based on crowdsourced data, real-time flood warnings and historical events, as well as river level information? The first research question was how to identify in timely manner at-risk locations of flooding using available national-scale real-time flood information such as USGS River Webcams, flood peak rates, and gauge height information sources across the United States. Nationally, there are an increasing number of near real-time, real-time and historical environmental data sources that can provide information on at-risk locations and river level information.

2. How can Rest/Streaming APIs and machine learning approaches help us to automate the spatiotemporal retrieval of data and gather the information from various web servers and applications? The second research question was to investigate the viability of automated retrieval of API data gathering, social media data collection (i.e. tweets geotagged by their account) in a timely and continuous way, based on multiple locations prioritized by potential risk of flooding.

3. Do constructed flood hydrographs over critical infrastructures provide valuable insight into flooding thresholds and frequencies? The third research question was whether construction of flood hydrographs over critical infrastructure such as roads and bridges help engineers in designing safe structures and in protection against economic losses due to maintenance of structures.
Our research aim was to develop and test a prototype that integrates historical and real-time solution based on these three research questions for intelligent assessment of at-risk locations and flooding impacts during a series of flooding events across South Carolina and North Carolina (the Carolinas) in the United States; that was also applicable to other countries and a range of weather-driven emergency situations.
CHAPTER 2

LITERATURE REVIEW AND RELATED RESEARCH WORKS

Big data analytics is where advanced analytic techniques such as IoT operate on Big data platforms (Russom, 2011). These techniques specially developed to analysis a large amount of data that may not be possible to understand by simply looking at the data sets. These approaches can be applied in many fields such as clustering, association analysis, prediction, etc.

Big data is the information asset characterized by high volume, velocity and variety that require a specific technology and analytical methods to translate its transformation into value (Mauro et al, 2016). Big data are described by three characteristic which are volume, velocity, and variety.

- Volume, the quantity of data, which is generated and stored. The key to identified Big data is the size of respected data.
- Velocity, the speed of data processing and generation method. Big data generally must be analyzed at a rate that matches the speed of data production.
- Variety, the types of data, which can be text, image, video, etc.

To analyze Big data, the manual approach of data collection would not be enough due to the size of data which inspired the creation of Hadoop. Hadoop is a framework designed specifically
for Big Data. Its primary component is Hadoop Distribution file System (HDFS), and MapReduce.

Pollard et al. (2018) investigated coastal flood risk management and emergency response procedures using five Big Data Approaches (BDAs): 1) Synthesis and harmonization of coastal datasets; 2) Handling and validating satellite data; 3) Integrating process-based modeling and Bayesian Networks; 4) Ensemble forecasting; and 5) Natural language processing of social media. Their research concluded that BDAs can be effective in the coastal flood risk management after overcoming the challenges of present BDAs. These challenges include whether BDAs can provide insights under conditions of incertitude and whether BDAs can be used by less tech-savvy researchers.

Pastor-Escuredo et al. (2018) analyzed different types of methods to evaluate Big data and addressed flooding impacts in real-time, flood segmentation and visualization, social media semantic and network analyses, and rainfall measurements. Their proposed framework underlined the design of real-time data structure provided by data-driven approaches. This work showed that for flood management, social media in combination with real-time presence data integrated with Call Detail Records could complement environmental warning mechanisms and provide an accurate proxy of flood warning.

Barker and Macleod (2019) developed a national-scale real-time Twitter data mining pipeline for the United Kingdom that collects social geodata from at-risk communities. Their tool pools national flood warning and river level Web data sources and computes at-risk locations to flooding in real-time.

Macchione et el. (2019) created a criterion for making a flow of information, using various non-conventional sources such as amateur videos, photo, topographic survey, news
reports, etc. The goal of this research was to show that the Big data can be used with a hydrodynamic model to numerically reconstruct flood dynamics across an urban setting. This analysis suggests great difficulties in performing flood analysis due to lack of conventional data related to hydrological and hydraulic variable, to perform 2-D numerical reconstructions and 3-D simulations.

Ning et el. (2019) implemented a prototype system to screen flooding photos from social media. This system designed with multiple key functions which are Tweets downloading, image downloading, flooding photo detection, and human verification via a WebGIS application. This work conducted a training dataset of 5,000 flooding photos that was built based on an iterative method; a convolutional neural network (CNN). In their research CNN was trained and applied to detect objects such as flooding in time lapse. They tested the trained CNN for Hurricane Harvey flooding at Houston, Texas as well as Hurricane Florence Flood (20180 in the coastal Carolinas. Their analysis revealed that CNN detected 95% of flooding in the Hurricane Harvey driven floods images whereas the model was able to detect 46% of flooding in Hurricane Florence driven flood images. Although, these results show some limitations of the trained CNN such as lower precision in detecting flooding object in social media images.
CHAPTER 3

METHODOLOGIES

3.1 DATA GATHERING AND ANALYTICS

This project deployed USGS historical and real time data and images as well as Twitter data. These data are publicly available and distributed across the nation. However, there is a need to automate identification and extraction of these massive Big databases and spatial information about flood events (Smith et al., 2015) particularly for real-time flood assessment. Accurately identifying localized flood extent is, however, difficult, especially during real-time particularly in mixed coastal and urban settings. The characterization of coastal urban catchment response during local flash flood events are shown to particularly benefit from Big data sources as well as social media feeds. Further, analysis of social geodata including Twitter feeds during flooding can identify geographical relations of at-risk locations, flooding, and vulnerable communities. Over past few years, several studies have focused on using social media data in a stand-alone manner for specific events and locations, although with no attempts to integrate these community datasets with real-time environmental data which are continually expanding and becoming accessible via public APIs. Current packages and web applications that harness flood related data have been generally developed for historical data collection and/or for specific emergency
events (Spielhofer et al., 2016). In other words, the existing tools have never monitored national level hazard events in real-time across USGS data networks as well as social media. Further none of the existing approaches evaluate the time-lapse flood images and engineering design metrics (i.e., flood frequency analysis; FFA). Here, we demonstrated and tested a national scale prototype flood data analytics pipeline, based on both historical and real-time flood warning and river level information, as well as crowdsourced processing of tweets. This included automated selection and analysis of large volumes of geotagged and relevant social media data, and recent advances in machine learning. Our developed machine learning approach trains custom image models with Google AutoML graphical interface then optimizes the model for accuracy, latency, and size.

3.2 **Data Analytics Approaches**

Our prototype is inspired by the impacts and damage of successive hurricane events in the southeast United States and by several discussions with emergency management officials. During several visits to South Carolina (SC) Emergency Office as well as virtual discussions with federal agencies such as USGS and Federal Emergency Management Agency (FEMA), we demonstrated the need to develop a prototype pipeline linking historical and real-time river level information, social media, and machine learning approaches to aid national level flood management and impacts assessment. Further, there is a need to automate Big data and crowd sourced information collection in real-time and create a map-based dashboard to better determine at-risk locations and flood situations across the United States. These needs and discussion along with deficiencies in existing data pipelines provided comprehensive roadmap tasks for this research.
FAIS, a national scale Big data pipeline, is smartly designed to poll flood related datasets from USGS and Twitter servers using APIs. The tool first gathers national flood peak rates and river gauge heights Web data sources in both historical and real time. Next, it intersects tweets with the USGS flood gauges to proximate at-risk areas to flooding-which tell us what areas will be flooded and what areas will be safe. The tool filters flood-relevant tweets with natural language processing using word embedding of tweets.

3.3 **MACHINE LEARNING AND IMAGE PROCESSING APPROACHES**

This study used Google Vision API and image processing approaches to detect objects in time lapse images. Google Vision API is an intelligent tool that detects and extracts information about objects and entities in an image, across a broad group of categories. This tool encapsulates machine learning models in an API that allows developers to use Computer Vision technology for classifying images into thousands of categories and assign them sensible labels and scores. Vision API detects objects in the images by following below functionalities

1. Detect and classify multiple objects including the location of each object within the image.

2. Build and deploy fast, high-accuracy models to classify images or detect objects at the edge, and trigger real-time actions based on local data.

3. Retailers can create an engaging mobile experience that enables the user to upload a photo of an item and immediately see a list of similar items

4. Detect text within images in more than 50 languages and various file types
5. Review images using Safe Search and estimate the likelihood that any given image includes adult content, violence, and more.

FAIS allows the user to use the Vision API directly or use AutoML Vision to train user custom machine learning model by simply uploading images and training custom image model. The application follows the following steps to detect the objects in the image using Google Vision API, notably

1. A user uploads an image from local storage to FAIS application
2. Once a new image arrives, FAIS requests the label detection service from Google Vision API.
3. Google Vision API utilizes machine learning tools to perform label detection on a request image.
4. Google Vision API sends the result back to the FAIS application.

The tool can detect individual objects and pieces of text and information within an image using two methods:

1. Using Vision API directly from the application to analyze images
2. Building custom models using the API to accommodate more flexibility for particular use case.

FAIS uses the first approach to directly communicate with the Google Vision API and detect the labels in images and score them. The application lets the user drag/upload an image in its interface and detect and recognize a bunch of information regarding the image. More specifically, FAIS uses Label Detection to annotate an image with a label (or "tag") based on the image content and name them. For example, a picture of a flooded road may produce a label of "flood", "road", or some other similar annotation. Label
Detection determines broader category contexts in different ways—for example, image labels such as “flood”, “water”, “river”, “floodplain”, etc. that cover broader categories of water resources objects.

3.4 Flood Frequency Analysis (FFA)

FAIS performs flood frequency analysis as a technique used by hydrologists to predict flow values corresponding to specific return periods or probabilities along a river. The tool uses “dataRetrieval” (DeCicco, 2018) and “xts” (Ryan, 2018) R libraries to retrieve annual peak flow rates for several years and calculates statistical information such as mean, standard deviation and skewness which are further used to create frequency distribution graphs. The tool currently fits Gumbel distribution to the annual maximum flood data and plots frequency curves. These graphs are then used to estimate the design flow values corresponding to specific return periods which can be used for designing structures such as dams, bridges, culverts, levees, highways, sewage disposal plants, waterworks and industrial buildings. Flood frequency estimates are useful in providing a measurement parameter to assess the damage corresponding to specific flows during flooding event. Along with civil infrastructure design, flood frequency estimates are also useful in flood insurance and flood zoning activities. Accurate estimation of flood frequency not only helps engineers in designing safe infrastructure but also in protection against economic losses due to maintenance of structures. It should be noted that the theoretical definition of return period is the inverse of the probability that an event will be exceeded in a given year. For example, a 10-year return period corresponds to a flood that an exceedance probability of 0.10 or a 10% chance that the flow will exceed in one year.
3.5 Crowdsourced Data Collection

Crowdsourcing citizen base data has increased its importance and popularity which can now apply in many research fields. There are multiple sources that contain crowdsourced data, one of the most up-to-date sources is social media. Social media is very well-known source of information for real-time flood studies, but also it can be misleading and unreliable. Social media platform allows the user to publish any story that they want but does not have automated reliable fact check software implemented to detect false information. Therefore, the user needs to be caution in the reliability and usage of these data as a proxy for any research activity.

Social media data is of interest to scientists as well as government and corporate interests. Many researchers recognize the importance of crowd sourced data, as a platform that intensifies commodification of Big data for large corporations and research activities. In other words, accessibility to voluntarily generated and often publicly published content on social networking and social media sites provides a strong draw for disaster related research. Social media, particularly Twitter, is increasingly used to improve situational awareness and two-way communication during hurricane and floods events (e.g., Kryvasheyeu et al., 2016). Stakeholders can inspect actionable information on real-time dashboards that integrate different sources and provide an information channel for the citizens and communities (Barker and Macleod, 2019).

Most of the social media platforms provided their own API to help access post. For example, Facebook provided Graph API which is a primary way to get data into and out of the Facebook platform. Instagram provides a very limited data access API on their site, but there exists an opensource API which provides flexible data access platform called
Instagram-API-Python. For Reddit, the user can use the Reddit Bot API to access the data. However, these social media platforms have their own strength and weakness. The reason why this study used Twitter is that it is less challenging to gather Twitter flow of information and Hashtag (#) driven topics for searching, although spatial locations and significance of natural catastrophes can influence the quality and amount of social geodata.

The interested data which can be analyzed from these sources can be images, videos, text, and flood depth (numeric) from volunteer inputs. Employing intelligent computing including machine learning allows alleviating information overload or data reliability issues. This provides crucial information to first responders and stakeholders during real-time event for map visualizations of flooded areas (Rao et al., 2017; Li et al., 2017). For time lapse data such as images and videos, a series of images can be analyzed over time using image processing techniques that include watershed analysis, image segmentation, gradient analysis, etc. In addition, these methods can compute flood characteristics from the images such as flood depth and inundation area. Twitter outcomes such as the sentiment analysis can also be performed for a tweet. This can be then used by social scientists to determine how do the citizens responded to the flooding event. There are several studies that showed crowd source data in conjunction with image processing are useful to identify at-risk locations and characterize flood depth and inundation areas such as the case studies for Sao Paulo, Brazil (de Assis et al., 2016), Jakarta, Indonesia (Elander et al., 2016), the River Elbe, in Germany (Herfort et al., 2014), and several urban locations in the United Kingdom (Barker and Macleod, 2019).
3.6 **Twitter APIs Tools**

This study first developed a python package targeting two sources of data i.e., USGS and Twitter. The package transferred to a web Python platform to collect the data during historical and real-time events and visualize flooding impacts. USGS collects and stores multiple water information data over time across the United States. They are river flow data (flood, streamflow, gauge heights), water quality, ground water levels, and precipitation at defined gauging stations which are strategically placed at the outlets of rivers and lakes. These placements allow USGS to correctly monitor and collect the data and compute several statistical indices related to the river flow across the nation. USGS provided two different types of flow data including real-time and historical records based on datetime.

The increasing interaction with social media and government agencies has shown that social media platforms are important tools to study floods. Twitter, one of the biggest social media platforms, contains many fast-paced real-time data and wide range historical information regarding news and events that occur in the local and global scales. Being the famous social media, local resident and government agencies use this platform to provide information during catastrophic events. For example, during Hurricane Florence, many residents in Columbia, SC tweeted the information about the local damage, road closure and shelter information. The government agencies such as National Weather Service (NWS), National Hurricane Center (NHC), South Carolina Department of Transportation (SCDOT), and USGS used Twitter to provided updates about the damaged infrastructure, emergency situations, and resources in real-time and during post-event. These data can be
combined with USGS records to identify at-risk locations and sending warning messages
to the citizens about flooding conditions.

FAIS application collects tweets using two different approaches which are Twitter
Search and Streaming APIs. These two methods are based on two used case of the
applications which are historical data gathering as well as real-time data collection.

3.6.1 Twitter Data Streaming

FAIS captures real-time tweets using Tweepy package. Tweepy is an easy-to-use
Python library for accessing the Twitter data. Twitter allows any authorized user to use
their API for free but with many limitations. Our developed application utilizes the search
service on twitter website that requires the users to request for developer account. The
developer account provides credential such as Access Token, Access Token Secret,
Consumer Key, and Consumer Secret to access Twitter functionality from outside of twitter
environment. To protect the credential, this project decided to create a Twitter Bot, and
deploy it at Heroku outside of the application access which can be controlled by the Heroku
User Interface. The administrator can choose to activate and deactivate the Bot and change
the keywords for Streaming services.

FAIS collects Twitter feeds in real time as well as during historical events and then
save filtered data to MongoDB cloud database. Due to the size of queried data, the filter
only keeps text, location, author, and date of tweets which are eliminated over 95% of
uninterested data. FAIS application has access to MongoDB cloud database without having
access to Twitter Bot. This enhances the privacy of user account information and data.
Figure 3.1 shows the workflow of Twitter Streaming approach designed within the FAIS pipeline.

![Figure 3.1—Tweet Streaming Relationship](image)

### 3.6.2 Development of Twitter Data Searching API

Twitter Search API is a functionality to search past or current tweets that has matches search criteria. Tweepy contains the search functionality but has a limited amount of search result, search frequency, and time constrained. To upgrade the functionality, the user requires to pay for the upgraded Twitter account. With the goals of keeping FAIS application free and open source, the alternative search method was implemented which was using the REST service freely from Twitter search function. Twitter search operation allows the users to look for tweets based on interested user account, keywords, language,
and time period but this is not an ideal procedure for the user to go through the website and gather all the interested tweets that they need. Therefore, this project used python “urllib”, standard library for Universal Resource Locator (URL) handling modules, to send specific request to Twitter search web service and gather the tweets. The URL can be then modified based on the user search criteria as shown in Figure 3.2.

```python
url = "https://twitter.com/i/search/timeline?f=tweets&q=%s&src=typd&max_position=%s"
url_get_data = ''
if hasattr(tweet_criteria, 'username'):
    url_get_data += ' from:' + tweet_criteria.username
if hasattr(tweet_criteria, 'since'):
    url_get_data += ' since:' + tweet_criteria.since
if hasattr(tweet_criteria, 'until'):
    url_get_data += ' until:' + tweet_criteria.until
if hasattr(tweet_criteria, 'querySearch'):
    url_get_data += ' ' + tweet_criteria.querySearch
if hasattr(tweet_criteria, 'lang'):
    url_lang = 'lang=' + tweet_criteria.lang + '&'
else:
    url_lang = ''
url = url % (urllib.parse.quote(url_get_data), url_lang, refresh_cursor)
```

**Figure 3.2**—Universal Resource Locator for sending request to Twitter in JSON format.

The gathering tool continuously sends request URL to Twitter and wait for the JSON response send back from Twitter. After receiving the JSON file, FAIS extracts interested tweets including username, text, retweets, likes, date, id. Permalink, user_id, media, URL, and the sentiment. The response JSON file contains a ‘min_position’ data file which can be used to move iteration forward. The iteration will keep running until the response exceeds maximum numbers of tweets limit set by the user. Note that the date limit is set to the request URL sent to Twitter.
3.7 USGS API for Data Gathering and Web Cameras

USGS provided many categories of information respected river system. FIAS application focus on gathering USGS discharge data (daily streamflow and flood data) as well as gauge heights, river web camera, Real-time Stream Data, and Historical Stream Data. The Stream Data is value of flow rate in cubic feet per second (cfs), and water level (gauge height) in feet.

3.7.1 USGS Historical Flood Data Collection

FAIS gathers USGS historical discharge data at national scale for any state, analyzes the information and plots the data. These data are recorded every 15 minutes to 1-hour intervals but update in different time steps across the stations. FAIS made USGS data collection seamless and straightforward by providing access to the station’s information in different including URL and ID name.

Gathers USGS historical data involved two steps include selecting the target state, the interested station, and date. After query criteria are entered, FAIS creates a request URL and sends it to USGS for collecting the data. The request URL is constructed as follow

```python
url = "https://nwis.waterdata.usgs.gov/"
url_criteria = ''
if hasattr(USGS_criteria, 'region'):
    url_criteria = USGS_criteria.region + '/nwis/uv'
else:
    url_criteria = 's/nwis/'
if hasattr(USGS_criteria, 'parameters'):
    for parameter in USGS_criteria.parameters:
        url_criteria += 'cb=' + parameter + '&'
url_criteria += 'format=rdb'
if hasattr(USGS_criteria, 'station_number'):
    url_criteria += 'site_no=' + USGS_criteria.station_numbers + '&'
if hasattr(USGS_criteria, 'since') and hasattr(USGS_criteria, 'until'):
    url_criteria += 'period=begin_date=' + str(USGS_criteria.begin) + '&end_date=' + str(USGS_criteria.end)
url = url_criteria
```

Figure 3.3—Request URL for USGS Historical Data
3.7.2 **USGS Real Time Flood Data Collection**

USGS monitors real-time flood data every 15 minutes to one-hour at state level. This interval creates a time lag that can be delayed but since the flood normally do not happen in 1-hour period the delayed can be endured.

FAIS application sends a URL request containing the region, state abbreviation, portion of the request URL as follow:

```python
data_url = "https://waterwatch.usgs.gov/waterdata/services/realtime?region=" + region + "&format=csv" usgs_csv = pd.read_csv(data_url, converters={'site_no': lambda x: str(x))
```

**Figure 3.4**—Request URL for USGS Real-time Data

The difficulty of this read_csv is that the station number needed to be recorded as a string to preserve the leading 0 on for station identification.

3.7.3 **River and Road Web Cameras**

In addition, FAIS gathers USGS Realtime cameras images and currently collects six Sevillian cameras images for South Carolina. These cameras are located at Rocky Creek near Wade Hampton, Rocky Branch at Whaley St., Columbia, PeeDee River near Florence, Lake Monltrie Trailrace Canal at Moncks Corner, SC, Tearcoat Brach at I-95 near Manning, SC, and Pocotaligo River at I-95, above Manning, SC. The image contains the current datetime of which it has been taken to help researcher categorize and view the data.

Department of Transportation in each state provides traffic data in real-time to the users through 511 web application. The application uses a dynamic mapping interface to
allow the user to select specific traffic camera and access the data in real-time. The cameras are strategically placed on the bridges and roads and along the interstate, allowing operators to continuously monitor road conditions. They also monitor rising and falling water stage over critical infrastructures such as roads and bridges. Currently FAIS gathers North Carolina (NC)’ 511 images in real-time. Several cameras that are in the coastal region of NC were selected from a larger set of devices to record roads flooding condition in real-time.

### 3.8 Web Development Framework

With the use of Python in the FAIS API tools, it was straightforward to use the same language and develop a web application. The seamless interaction between the API and Python motivated this study to use Django as a web framework that encourages rapid development, clean, and pragmatic design (Django Software Foundation, djangoproject.com). The key point of making it an ideal development framework is due to its focus on automation as much as possible and adhering to the “don’t repeat yourself (DRY)” principle. It allows the user to develop a web application using Python element along with a classic web development language like HTML, JavaScript, jQuery, etc. This means that developers are not stuck with limited approaches of solving the problem. This interaction can also be a downside of this development frameworks as it requires the knowledge of different languages to debug like Ruby on Rails. (Julia et el, 2009) conducted a study comparing Django development framework with Ruby on Rails, and CakePHP using five different evaluations criteria which are user interface development, maintainability, data management and migration, testability, and popularity. The result
shows that Django perform best on those scores, but other tools is performed in the similar result range.

Django provides the widest range of libraries for rapid and pragmatic design including Twitter APIs and machine learning. FAIS is version-controlled using GitHub since Git is a widely used version control system and a popular repository hosting service for free and open source.

Figure 3.3 shows the overall workflow and FAIS pipeline architecture for data retrieval. As illustrated, FAIS combines multiple APIs with image processing algorithms and FFA to provide both historical and real-time information about flood-risk incidents. Our developed prototype offers an end-to-end, open source, web-based, pipeline architecture to address the crucial issue of how first responders and decision makers can be better informed in emergency management situations. To achieve our aim and answer three research questions, we tested FAIS operationally during Hurricane Dorian flooding event across the Carolinas between September 04–06th, 2019. In next sections, we present a systematic description of the pipeline steps and major design options and outcomes.

Figure 3.5—FAIS workflow and structure
CHAPTER 4

DISCUSSION

4.1 APPLYING AND TESTING FAIS DURING REAL-TIME AND HISTORICAL FLOODING EVENTS

There were widespread flooding events during January to December 2019 in the Carolinas, when the prototype was Beta tested. These events caused localized and major flooding as well as above average river flows for most of river systems in the Carolinas. FAIS was assembled and tested in real time during September 04-08, 2019 when Hurricane Dorian flooding caused damages in the Southeast US. The prototype was then applied nationally across US, running for both single flood event as well as continuous flow data collection.

4.2 DESCRIPTION OF PIPELINE STEPS AND DESIGN OPTIONS

This section focuses on addressing three research questions stated above. We systematically describe a series of steps and major design options and decisions that enabled production of a national scale prototype machine learning and data gathering pipeline, based on historical and real-time flood warnings and river level information.
4.2.1 RETRIEVAL OF REAL-TIME SOCIAL GEODATA USING TWITTER APIs

This section focuses on designing and accessing Twitter data. This study used Tweepy library as a Search API and developed a Streaming API using Twitter Bot. FAIS includes two options for Twitter data gathering i.e., the Search and the Streaming APIs. Search API was more suited to singular and specific queries for tweets, whereas the Streaming API provided a real-time stream of tweets. FAIS pipeline uses both Search and Streaming APIs to highlight both historical and real-time information on flooding events in the communities. Implementation of Twitter Search and Streaming APIs was done during Hurricane Dorian event (September 04-06, 2019) as well as during historical events (Hurricanes Matthew and Florence) in the Carolinas. An example from each event is presented with intersecting at-risk locations and geotagged tweets. A web-based console and a visualization tool- GeoJSONLint13 were used to view results and inspect the polygons.

We monitored georeferenced tweets, filtered by keywords and queries across the shortlisted areas (the Carolinas) during Hurricane Dorian, which retrieved tweets whose geotag intersected: a maximum of up to 12 locations bounding boxes were permitted at a time. These locations were monitored during Hurricane Dorian event for any updated tweets whose geotag intersected a bounding box, which constituted a “match”. Location queries used with Twitter's Streaming API and tweets’ place with GPS coordinates were then intersected with watershed polygon to determine the prioritized areas to flooding. The retrieved tweets were then stored in a MongoDB database which is widely used as an ideal open-source document database to store JSON format files.
FAIS prototype was in operational use between 04 to 06 September 2019 and constantly ran the Twitter Bot to determine at-risk locations over time. The stored tweets provided a real-time dataset which then explored and used to prioritize at-risk watershed systems and location for Dorian flood simulation. This provided a general indication of the proportion of potentially relevant tweets that can be used to identify flooded areas and improve flood situational awareness. Data retrieval of Dorian flooding was explored via basic text query searches such as “Hurricane Dorian Flood” and centroids were added to tweet’s geotag for creation of map-based visualizations.

4.2.2 Identifying At-risk Locations to Flooding Events

This section addresses the first research question: How to programmatically identify at a national-scale (across the United States) areas at-risk of flooding based on crowdsourced data, real-time flood warnings and historical events, as well as river level information? To address this question, a Python code was first developed that ran on a local computer server, the script gathers real time flood data as well as Twitter feeds in order to identify areas at-risk of flooding from the latest USGS data sources. A period of 6 to 12 hours was chosen as an intended trade-off between tracking the latest at-risk area forecasts (API updates varied between 15 min and 6 hours depending on flood data time step), and that allowed capturing reaction from those at-risk areas on Twitter.
4.3 Applying and Testing FAIS Pipeline in Real-time During Hurricane Dorian

To identify at-risk locations in real-time, a shell script in Python runs on a local computer server, the script is reset every 3 h in order to update areas at-risk of flooding from the latest national and environmental data sources as well as Twitter feeds. A period of 15 minutes was initially chosen as intended trade-off between tracking the latest at-risk area forecasts. API updates can be varying between 15 minutes to several hours but based on real-time testing the period extend to 3 hours to allow some reaction time from those areas on Twitter. However, the choice of time period depends solely on the project requirements as well as the impacts of flooding events. FAIS first cycled through a set of USGS web addresses for river gauge height readings, parsing these flat files using Python web scraping technique and obtained all the latest river levels. Each river level reading was compared with its respective long term cached average level, to identify the highest relative river levels in real-time. The highest river level then intersects with watershed polygons as well as geotagged tweets to identify at-risk locations to flooding. Geotagged tweets coordinates considered as a center-point for approximately 16 km wide square boxes. The morphological characteristics of these hot spot areas were then explored to understand the type of watershed (rural vs. urban) and runoff generation and processes at the watershed scale. Delineating at-risk areas can be further improved by populating critical infrastructure and flood defense structures. The prototype programmatically identifies at national-scale (across the United States) at-risk areas of flooding in real-time.
4.4 Applying and Testing FAIS Pipeline During Historical Events

In this section, we successfully identified a dynamic set of at-risk areas using Twitter Streaming approach for Hurricane Florence event. At-risk areas to flooding were identified by intersecting the geotagged tweets with watershed polygons and river gage heights. The at-risk locations were identified based on the rules set out in real-time event. The users have the flexibility to change the keywords and Twitter Streaming Bot filters the tweets and saves the data in MongoDB database.
4.5 IMAGE ANALYSIS USING GOOGLE VISION API AND IMAGE PROCESSING

ALGORITHM

FAIS uses Google Vision API for object detection of time-lapse flooded images. Google API first trains the automated ML Vision models and then labels the datasets. This provides custom label detection with scores. We used flooded and non-flooded images of Hurricane Florence driven floods to detect the objects in the images. Figure 4.3 shows the detected objects including flood (91%), land lot (91%), and Asphalt (89%) in flooded image while the major label and score are given to sport venue (89%) and residential area (85%) in non-flooded image. Google Vision API detected the objects in flooded bridge before and after Hurricane Florence flooding event (see Figure 4.4). A number associated with every returned label annotation, representing the Vision API's assessment of the label's accuracy. Confidence scores range from 0 (no confidence) to 1 (very high confidence).
FAIS results were also evaluated using various flooded images in US such as Iowa floods 2019, Hurricane Harvey driven floods for Texas 2017, and Hurricane Irma driven floods in Florida 2017. These results proposed the superiority of Google Vision API for identifying flood hazard areas from image sequences. This information is critical for flood emergency management. Indeed, the capability to detect temporal changes in image sequences is crucial for an automated image-based flood alarm system and disaster-monitoring applications. This image-based flood warning information can facilitate proactive monitoring and assessment, and early warning to rising water levels and associated inundation in real time. This information can also help with development of an accurate flood forecasting system and inundation modeling for those at-risk areas identified during real-time.

Figure 4.3—Flooded (right) and non-flooded (left) images with detected labels and scores during Hurricane Florence event at New Bern, NC.

Figure 4.4—Flooded (right) and non-flooded (left) images with detected labels (bridge, river, water, etc.) and scores during Hurricane Florence event at New Bern, NC.
4.6 USGS Data Collection

USGS collected data include the date, discharge (cfs), and gage height (ft). Figure 4.5 shows USGS historical data including gauge height average and discharge during August to October 2019 at Washington Creek at Windigo (station id 04001000), Michigan.

Figure 4.5—Historical Data of USGS 04001000, in Michigan, from 08/01/2019 to 10/31/2019

The real-time data as of the October 18, 2019 at 11:53 am is gathered at the state of North Carolina. There are 291 real-time gauging stations and FAIS combined these real-time data with geotagged tweets to identify at-risk locations to flooding.
Another example of real-time data gathering as of the October 18, 2019 at 11:53 Am for South Carolina. There are 107 real-time gauging stations

<table>
<thead>
<tr>
<th>Station</th>
<th>ID</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Flow (cubic ft)</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRIAR CR Above Colony Rd At Charlotte, NC</td>
<td>0214649022</td>
<td>35.1753</td>
<td>-80.8308</td>
<td>3.04</td>
<td>2.14</td>
</tr>
<tr>
<td>BRIAR CR At Independence Blvd At Charlotte, NC</td>
<td>0214643770</td>
<td>35.2108</td>
<td>-80.8008</td>
<td>0.0</td>
<td>0.72</td>
</tr>
<tr>
<td>BRIAR CR #1, Edward Branch Near Charlotte, NC</td>
<td>0214643060</td>
<td>35.2061</td>
<td>-80.8056</td>
<td>0.0</td>
<td>1.17</td>
</tr>
<tr>
<td>BRIAR CREEK At Providence Rd At Charlotte, NC</td>
<td>021464449</td>
<td>35.1825</td>
<td>-80.02</td>
<td>0.0</td>
<td>1.14</td>
</tr>
<tr>
<td>BRIAR CREEK Near Charlotte, NC</td>
<td>0214645325</td>
<td>35.2361</td>
<td>-80.7711</td>
<td>0.07</td>
<td>1.25</td>
</tr>
<tr>
<td>CAMPBELL CREEK NR Charlotte, NC</td>
<td>021465562</td>
<td>35.1867</td>
<td>-80.7367</td>
<td>0.22</td>
<td>0.98</td>
</tr>
<tr>
<td>CUPPET CREEK NR Charlotte, NC</td>
<td>021463348</td>
<td>35.1408</td>
<td>-80.8289</td>
<td>1.15</td>
<td>2.18</td>
</tr>
<tr>
<td>EDWARDS BRANCH AT SHEFFIELD DR AT CHARLOTTE, NC</td>
<td>0214643825</td>
<td>35.2031</td>
<td>-80.7722</td>
<td>0.08</td>
<td>0.72</td>
</tr>
<tr>
<td>IRVINS CREEK AT SR3168 NR CHARLOTTE, NC</td>
<td>0214657975</td>
<td>35.1598</td>
<td>-80.7153</td>
<td>0.25</td>
<td>0.84</td>
</tr>
<tr>
<td>IRWAN CR At Statesville Ave At Charlotte, NC</td>
<td>02146111</td>
<td>35.2619</td>
<td>-80.8369</td>
<td>1.1</td>
<td>1.85</td>
</tr>
</tbody>
</table>

**Figure 4.6**— North Carolina real-time flood data as of the October 18, 2019

**Figure 4.7**— South Carolina Realtime flood data as of the October 18, 2019

### 4.7 Flood Frequency Analysis (FFA)

FAIS provides FFA to estimate extreme flood quantiles that combines elements of observational analysis, stochastic probability distribution and design return periods. FAIS currently uses Gumble distribution to compute FFA for any given flood gaging station in
US. Figure 4.9 shows FFA for USGS 02172035 TURKEY CREEK ABOVE HUGER, SC. As illustrated annual flood peak of ~6000 cfs represents a design return period of 15-year. Although, the resulted FFA curve may vary using different probability distribution such as Peason type III, Gamma, and Normal distributions.

![Flood frequency analysis for USGS 02172035](image)

**Figure 4.8**—Flood frequency analysis for USGS 02172035

USGS 02172035 gaging station located at in the coastal plain SC. FFA for this station proved that frequent flood events make critical contributions to the upper tail of flood peak distributions. However, hurricane driven floods trigger recorded annual flood peaks in this region. Coupling FFA with FAIS application is crucial for design flood analysis, but the frequency of flood event may be under sampled in USGS station thus poorly served by design storm assumptions. Therefore, care should be given when the outcomes use for any structure design purposes.
4.8 LIMITATIONS OF OUR PROTOTYPE AND FURTHER WORK

FAIS prototype as a national-scale Big data analytics pipeline could be improved further in several ways. In the FFA section, we highlighted revising civil infrastructure design metrics based on nonstationary approaches that can be integrated into the FAIS application. In addition, the size of bounding boxes (currently 16 km from the centroid) require further attention and evaluation, given the boundary effect on tweets intersect bounding boxes and yielding a match. In an urban setting, increasing the size of bounding boxes would likely include significantly more tweets thereby more accuracy on the chosen proximity. The selected tweets can be also overlaid with other data layers such as census data, population density, rainfall radar data, and Federal Emergency Management Agency (FEMA) flood hazard map. This information can be further used by stakeholders to evaluate flood situational awareness and emergency management.

FAIS application provides at-risk areas assessment over the entire process of flood disaster management. For future development, a linking approach can be proposed to increase system autonomy during flood risk assessment and data retrieval tasks. The system can further combine combines many data sources such as National Hurricane Center data and National Weather Service forecast along with geospatial data, attribute and statistical aspects and describe them in the form of Big data analytics and assessment. FAIS workflow evaluation and processes can be automated to effectively reduce man-machine interactions and the dependence on human experience. This will help satisfy the need for timeliness and reliability during emergency responses and management.
CHAPTER 5

CONCLUSION

This study demonstrated a Big data analytics prototype that combined Internet of Things (API) and machine learning to integrate historical and real-time environmental data (river levels and flood warnings) at the national-scale and advanced at-risk areas identification using flood relevant tweets. FAIS was Beta tested operationally for 2 days period of Hurricane Dorian flooding across the Carolinas with over 15,000 geotagged tweets collected from 38 dynamic, potentially at-risk areas. In this study, we demonstrated an adaptable and flexible solution for how to seamlessly integrate open-source libraries, in a continuous and timely manner for real-time flood risk assessment. We acknowledge that there is a need to evaluate FAIS pipeline with stakeholders and first responders. In addition, there is a potential to apply FAIS application in other states and in other emergency situations.

FUNDING SOURCES

This research was supported by the National Science Foundation's (NSF) Directorate for Engineering (ENG) (Grant # 15520-FB61).
REFERENCES


[Li, Caragea, Caragea 2017] Li H., Caragea D., Caragea C. Towards practical usage of a domain adaptation algorithm in the early hours of a disaster Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management (2017), pp. 692-704

Environmental model access and interoperability: the GEO model web initiative.
Environmental modelling & software. v. 39, pp. 214-228.


[Nativi, Mazzetti, Geller 2013] Plekhanova, Julia, Evaluating web development frameworks: Django, Ruby on Rails and CakePHP. Institute for Business and Information Technology.


[Spielhofer, Greenlaw, Markham, Hahne 2016] Spielhofer, Thomas, R. Greenlaw, D. Markham, A. Hahne. Data mining Twitter during the UK floods: investigating the potential use of social media in emergency management Information and Communication Technologies for Disaster Management (ICT-DM), 2016 3rd International Conference on (pp. 1-6), IEEE (2016)