Information Processing and Management 59 (2022) 102735

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A Natural Language Processing Approach to Understanding Context in the Extraction and GeoCoding of Historical Floods, Storms, and Adaptation Measures

Kelvin Lai [a], Jeremy R. Porter [a,b,*], Mike Amodeo [a], David Miller [a], Michael Marston [a], Saman Armal [a]

[a] First Street Foundation, 215 Plymouth St., Brooklyn, NY 11201
[b] City University of New York, 365 5th Ave., New York, NY 10016

ARTICLE INFO

Keywords:
Information Extraction
Natural language processing
Floods
Machine Learning
Newspapers

ABSTRACT

Despite the known financial, economical, and humanitarian impacts of hurricanes and the floods that follow, datasets consisting of flood and flood risk reduction projects are either small in scope, lack in details, or held privately by commercial holders. However, with the amount of online data growing exponentially, we have seen a rise of information extraction techniques on unstructured text to drive insights. On one hand, social media in particular has seen a tremendous increase in popularity. On the other hand, despite this popularity, social media has proven to be unreliable and difficult to extract full information from. In contrast, online newspapers are often vetted by a journalist, and consist of more fine details. As a result, in this paper we leverage Natural Language Processing (NLP) to create a hybrid Named-Entity Recognition (NER) model that employs a domain-specific machine learning model, linguistic features, and rule-based matching to extract information from newspapers. To the knowledge of the authors, this model is the first of its kind to extract detailed flooding information and risk reduction projects over the entire contiguous United States. The approach used in this paper expands upon previous similar works by widening the geographical location and applying techniques to extract information over large documents, with minimal accuracy loss from the previous methods. Specifically, our model is able to extract information such as street closures, project costs, and metrics. Our validation indicates an F1 score of 72.13% for the NER model entity extraction, a binary classification location filter with a score of 73%, and an overall performance only 8.4% lower than a human validator against a gold-standard. Through this process, we find the location of 27,444 streets, 181,076 flood risk reduction projects, and 435,353 storm locations throughout the United States in the past two decades.

1. Introduction

Floods are the most frequent natural disasters, causing serious economic damage and human morbidity (Morss, 2010). In the United States alone, flooding is responsible for 99 lost lives a year in a 10-year average (US Department of Commerce, 2019a), 8250

* Corresponding author.
E-mail address: jp3323@columbia.edu (J.R. Porter).

https://doi.org/10.1016/j.ipm.2021.102735
Received 28 June 2021; Accepted 29 August 2021
Available online 9 October 2021

deaths in the last 80 years, and accounts for about 22.5% of mortality rate due to the weather (US Department of Commerce, 2019b). Over the last three decades, the U.S. has sustained more than 250 weather and climate disasters, with a cumulative cost of $1.75 trillion. About 80% of this loss is associated with floods, storms and tropical cyclones (Smith, 2020). Despite the financial, economical, and humanitarian impacts of flooding, nation-wide data on historic weather and climate data is typically limited to government sources such as the National Oceanic and Atmospheric Administration (NOAA), the United States Geological Survey (USGS), and the Federal Emergency Management Agency (FEMA), or held commercially by private companies. Even so, data collected by both government and commercial databases have issues. On one hand, government sources often cover a wider geographical location than at the parcel level, many river and coastal areas have yet to be mapped (ASFPM, 2020), and even areas that are mapped have limitations and underestimate risk for populations (Technical Mapping Advisory Council, 2015). While on the other hand, private databases are difficult to access freely, contain reporting errors and biases (Kron et al., 2012), and fail to detect a large number of flood events (de Bruijn et al., 2019). As a result of these inadequacies, it becomes extremely difficult for individuals to understand if they are in a flood area that puts them at risk of monetary or health damages, local communities performing informed decision making, and for external researchers to obtain an information-rich dataset for further research into flood patterns and protection.

Over the past two decades, accessibility to online resources has created a plethora of data available, and provided an opportunity to compensate for the scarcity of authoritative data resources. In the more recent years, special attention has been paid to the role of big data in natural disaster management. Be in like manner, this study relies on online resources to harness the power of text mining and create a flood database for the contiguous United States. Our approach is based on named entity extraction on publicly available articles using big data analytics.

In the context of natural disaster management, many studies (Kaufhold, Bayer, & Reuter, 2020; Ghafarian & Yazdi, 2020; Kozlowski et al., 2020; Dutt, Basu, Ghosh, & Ghosh, 2019; de Bruijn et al., 2019; Liu, Kar, Montiel Ishino, Zhang, & Williams, 2020) focused on social media as a preliminary source of big data. The high degree of visibility, real-time updates, and immense user presence, persuaded researchers to leverage crowdsourced data to identify real-time flood events and improve disaster communication during an event. However, despite all the benefits, the reliability of social media data poses a large issue in extracting accurate flood information. The pervasiveness of social media leaves it prone to the spread of false information (Oh, Kwon, & Rao, 2010) which in many cases are more likely to receive more attention than accredited news (Ortiz-Martínez & Jiménez-Arcia, 2017). Furthermore, the propensity for more newsworthy events, lack of accuracy in the latitude/longitude coordinates, and the inconsistent and subjective reporting of events are some other factors that compromise the quality of crowdsourced data (Arthur, Boulton, Shotton, & Williams, 2018). Besides the reliability issues, structural limitations such as the maximum character number, lack of context and shorthand notations and abbreviations impede on benefits of using the data (Karimzadeh, Pezanowski, MacEachren, & Wallgrün, 2019)bib18.

Social media however is not the only online medium to become easily accessible. Online newspapers have also become widespread, and contain hundreds of years of rich cultural, social, and historical information (Yzaguirre, Smit, & Warren, 2016). In particular, papers from local communities often provide detailed location-specific accounts of flooding and storms such as flooded roads and damages, or proposals and updates for flood risk reduction projects built in response to the threat of future environmental damages. With our goal of compiling the locations of historical floods and risk reduction projects across the United States, these news articles are an invaluable source of data. One of the main challenges in the creation of a nation-wide flooding dataset is the difference in newspaper writing structures, reporting style of flood details from a local newspaper compared to a national newspaper, and the reliability of information from different newspaper publications. Researchers in the past focused on extracting data from a single geographical location and publishers such as The Chronicle Herald from Nova Scotia, Canada (Yzaguirre, Smit, & Warren, 2016) and the Montreal Gazette from Montreal (Zarei & Nik-Bakht, 2019) to circumvent this issue. However, we cannot apply the same circumvention method since we require our model to extract flood data from any newspaper in the United States. Researchers have also recently applied techniques to extract information from complex multi-paragraph documents such as legal documents (Ji, Tao, Fei, & Ren, 2020), interpret rich banking documents (Oral, Emekligil, Arslan, & Eryiğit, 2020), and merge multiple streams of information to extract crime data from newspapers (K & Thilagam, 2019). Using similar techniques we leverage information extraction, geo-spatial data, and entity-based relationship graphs to create a detailed historic flood and risk reduction project dataset over the Contiguous United States based on online newspapers.

In that regard, the main contributions of this paper is four-fold:

- We created a hybrid named entity recognition model using rule-based and deep learning approaches to extract named entities from geographically-wide, multi-paragraph documents, and extract related entities
- The use of syntactic neighbors instead of the closest neighbour heuristic approach to extract entity relationships of flooding and toponym entities. To the best of the author's knowledge, this is the first of its kind to apply syntactic neighbors to perform such tasks
- We created a detailed historic flood dataset that covers the contiguous United States over the last two decades. This dataset includes flood event information such as the type, severity, and location of these events, opening a new dataset for further flood pattern analysis
- To the knowledge of the authors, we created the first flood risk reduction project dataset that covers the contiguous United States. While the US Army Corps of Engineers (USACE) have a National Levee Database (NLD), there are a large inventory of levees outside of USACE's authority, and the condition of the nation's levees is largely unknown (ASCE 2017) Infrastructure Report Card, 2017

1.1. Research objective

The present study pursue the following research objectives:

- The existing named flood entity extraction models typically only use a rule-based approach (Yzaguirre, Smit, & Warren, 2016), or only a statistical approach (Zarei & Nik-Bakht, 2019). Although these approaches work well for similarly structured documents and over small geographic areas, we investigate the effectiveness of a hybrid rule-based and machine learning approach to recognize named flood entities over a large variety of documents that differ in length, structure, and content with minimal accuracy loss.
- Previous research has focused on identifying the existence of a flood through the use of big data in social media. However, newspapers often contain additional information on floods such as the cause, intensity, and a precise location of the flood event. We investigate the practicality of extracting additional named entities, in addition to just the "flood" keyword.
- Following from the extraction of additional named entities, we investigate the usage of recent techniques in information extraction to semantically link different flood detail entities with their associated flood events, and compare this to the closest neighbor heuristic.

The remainder of the paper is structured as follows: Section 2 examines the Background and Literature Review. The problem is defined in Section 3. Section 4 explains our methodology, experimental dataset, and metrics. Finally, Section 5 analyzes the results of the methodology followed by the conclusion in Section 6.

2. Background and Literature Review

This section provides background information on existing flood detection and flood information models, as well as background on more recent information extraction and named entity recognition models

2.1. Flood Extraction

In the specific domain of information extraction in flooding data, there has been much research in real-time flood detection using social media (Ahmad et al., 2019; de Bruijn et al., 2019; Liu, Kar, Montiel Ishino, Zhang, & Williams, 2020; Moore and Obradovich, 2020; Rossi et al., 2018; Smith, Liang, James, & Lin, 2017; Wang, Mao, Wang, Rae, & Shaw, 2018). However, as noted before, social media has issues in quality (Arthur et al., 2018) and content (Karimzadeh, Pezanowski, MacEachren, & Wallgrün, 2019) due to abbreviations and incomplete word spellings, and is thus difficult to extract the same detail of information as online newspapers. Furthermore, many of the techniques used in these papers are not as easily applicable, such as extracting a location from a user's profile. Meanwhile, other researchers used satellite imagery to accurately identify the presence of a flood (Ahmad et al., 2019; Bischke et al., 2017). However, gathering high resolution satellite images of the contiguous United States for an extensive period of time would be extremely difficult and time consuming to process.

For flood information extraction using online newspapers, previous research has focused primarily on newspapers from a single source, and also only on the existence of a flood. Zarei & Nik-Bakht's research in 2019 on automatically identifying flooding focused on newspaper and news websites from Montreal, Canada, while Yzaguirre, Smit, & Warren's text mining on newspapers in 2016 focused on a major regional newspaper in Nova Scotia, Canada. In either case, with their model only spanning a small geographical area, a single news source, and not extracting additional information of the flood such as the reason or intensity, it is difficult to scale their work to a nation-wide model for the United States while maintaining accuracy and detail.

2.2. Named Entity Recognition, Relationship Extraction, Entity Linking, and Geocoding

Information extraction (IE) is defined as the task of automatically extracting structured information from unstructured and/or semi-structured text (Oral, Emekligil, Arslan, & Eryiğit, 2020), while the goal of named entity recognition (NER), a subtask of information extraction, seeks to label names of people, places, organizations, and other entities of interest in text documents. The three major approaches to NER are lexicon-based, rule-based, and machine learning based (Gudivada & Arbabifard, 2018). With the focus on the latter two, each approach has pros and cons. A rule-based model approach generally requires minimal labeling of training data, but also requires clear patterns to feature engineer the rule around, and extracts the typical and common occurrences of the entity. In contrast, the machine learning model requires less clearly structured patterns and is able to generalize to a wider variety of entity occurrences, extracting entities based on the clues of the local context. However, this approach requires much more training data. Related to NER and machine learning approaches, domain adaptation (Kouw and Loog, 2019) is the use of a previously trained separate model, modelled on a different dataset and target, to instantiate the training of another model. The benefits of domain adaptation and leveraging the learned embeddings include minimizing the impact of sample bias by adding generalizability and improving the NER accuracy due to faster convergence. Overall, while hybrid NER models (use of two or more approaches) have been used in other areas of interest (Gabbard, DeYoung, Lignos, Freedman, & Weischedel, 2018, Dias, Boné, Ferreira, Ribeiro, & Maia, 2020, Ruokolainen, Kauppinen, Silfverberg, & Lindén, 2020), this approach has yet to be applied to the area of flood extraction, and the techniques and application of NER in the extraction of entities from textual sources has grown in recent years, resulting in an opportunity to leverage similar techniques to the flooding domain to achieve our research objectives.

Beyond extracting entities using NER, the difficulty in extracting information from larger texts is that the information to extract

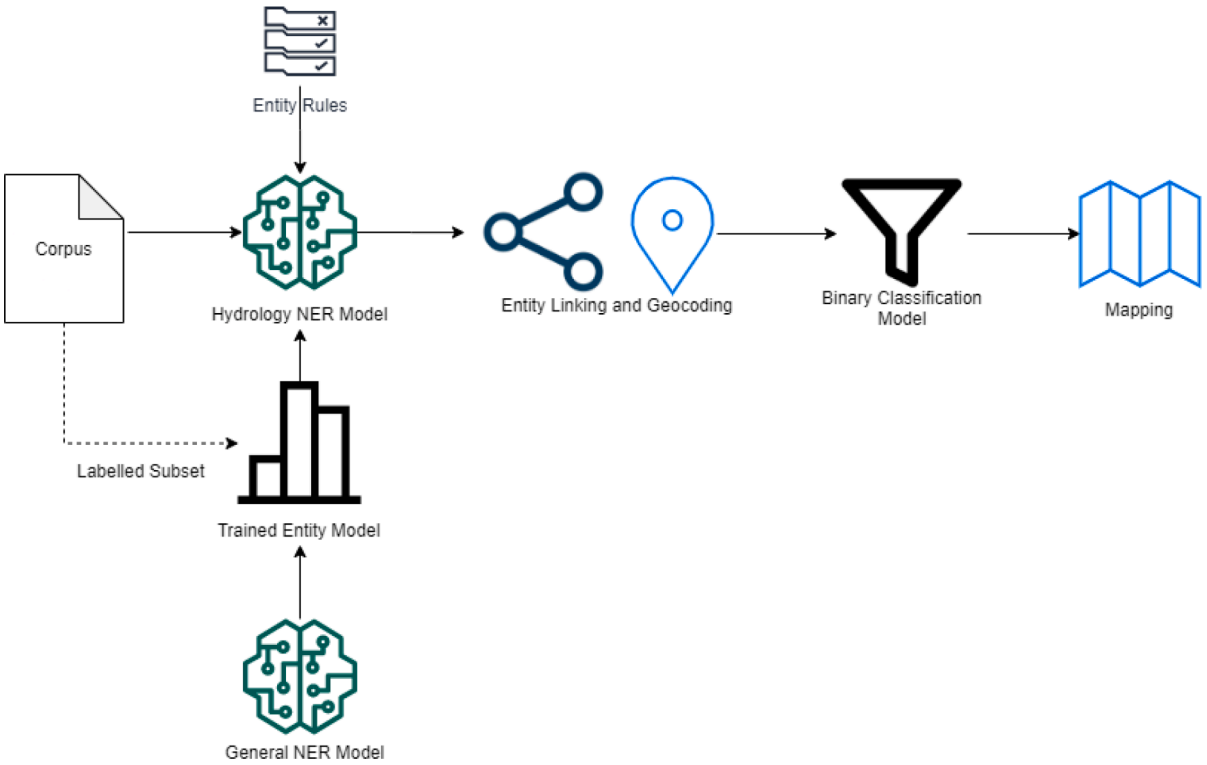

Fig. 1. Proposed High-level Full Project Process.

may be spread over multiple sentences and paragraphs. To handle this issue, the application of a relationship extraction (RE) is used to pair extracted entities, and entity linking (EL) is used to resolve ambiguous entities. Relationship extraction is the task of extracting semantic relationships from a text and usually occurs between two or more entities of a certain type (NLP-progress, 2021a). An example of RE is "There were apples, pears and oranges in the bowl.", in which we can extract the relationship between the pears and the bowl as "in''. Entity linking (EL) in contrast is the task of recognizing and disambiguating named entities to a knowledge base (NLP-progress, 2021b), such as correctly resolving ''Toronto, Ohio'' to a city in Jefferson County, Ohio, US, instead of ''Toronto, Ontario, Canada''. With this vein of work, researchers have made successful advancements in the extraction of information from complex documents. For example, in 2019 K and Thilagam created a crime base by applying an NER model to extract entities from online newspapers, with a similar structure of the knowledge base as our dependency parser graph. Others expanded on the research of EL such as (Munnelly and Lawless, 2018) work to create a knowledge base on Irish biographies, and Gupta, Banerjee, and Rubin work in 2018 on mammography reports. However, to the best of the authors knowledge, the use of a dependency parser, RE, and EL to extract flood information has yet to be done.

Finally, to extract a coordinate from an address, geocoding or geoparsing is used to convert extracted addresses and locations to a geographic coordinate by comparing the address to a knowledge base. For example, Yzaguirre, Smit, & Warren used GeoNames.org's gazetteer to geocode locations in 2016. Similarly, in 2019 Karimzadeh, Pezanowski, MacEachren, & Wallgrün geocoded locations from unstructured text tweets using NER algorithms and a GeoNames.org's gazetteer as well. Others used Google's geocoding API to get coordinates of street addresses (Laumer et al., 2020). Overall, while not a novel step, the use of a geocoder is important in both obtaining spatial information and performing entity linking and entity disambiguation.

3. Problem definition

To achieve the research objectives of this paper we look at the use of a hybrid NER model on online newspaper articles to create a nation-wide dataset with location-specific flood and risk reduction project information. Specifically, we want this model to be generic and able to capture as much information as possible since it needs to be applied to a wide variety of newspaper structures and writing styles. We also want the process to link extracted entities from news articles, allowing us to identify related and unrelated entities in the article. Finally, we also need to distinguish and filter locations of floods and flood reduction projects, and do this through the use of a confidence value for every extracted location.

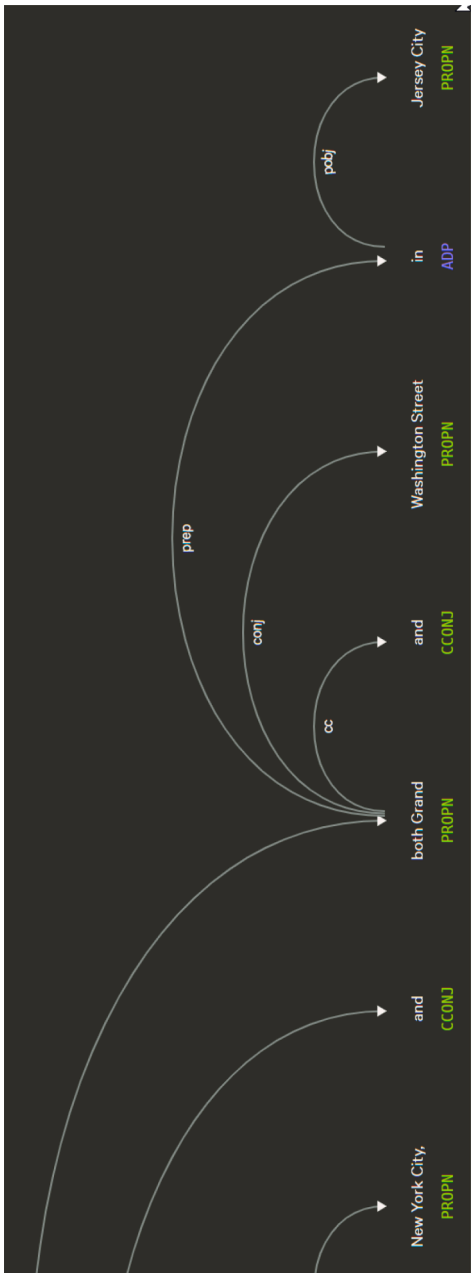

Fig. 2. An example of applying spaCy's syntactic neighbors to the sample sentence.

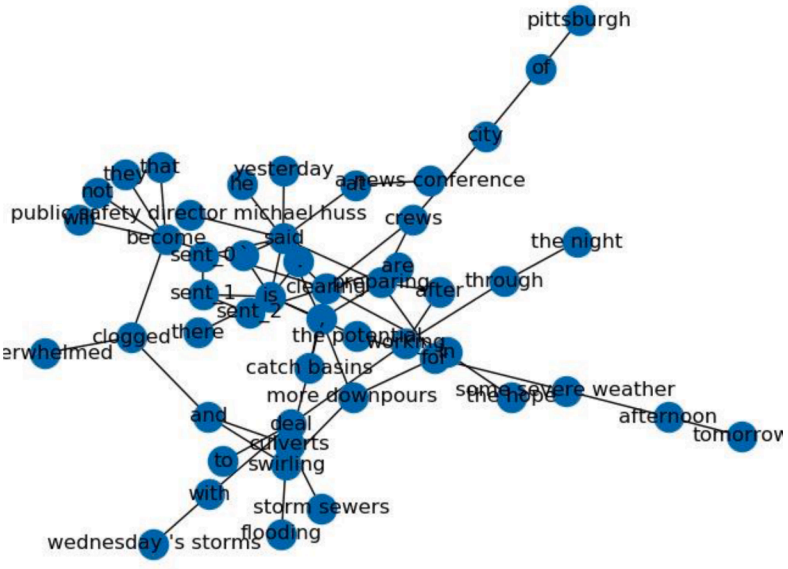

Fig. 3. Graph of Syntactic Neighbors for a Sample Paragraph.

4. Methodology

To achieve our goal, the proposed system can be broadly seen in Fig. 1 and consists of three main steps. 1) information extraction through the use of a hydrology NER model, 2) relationship extraction, entity linking, and geocoding locations, and 3) location filtering through the use of a binary classification model. The following subsections describe these individual steps in detail, while Section 4.4 describes the training dataset and evaluation metrics used in the model.

4.1. NER Model

Corresponding to the first step with the creation of the hydrology NER model in Fig. 1, using the (spaCy, 2021a) library, we created a custom NER model to tag words within the text with an entity from a predefined list such as hurricanes and storms. The custom model uses a hybrid rule-based and machine learning model for running NER. The entity rules used in the hydrology NER model are created by identifying common structured patterns in the corpus for the entities, such as street names where the entity consists of an optional adjective and/or noun followed by a street suffix like 'Street'. Since many roads consist of a similar pattern, we encode this rule into the model and repeat this process for other entities such as hurricanes, storms, flood risk reduction projects, and bodies of water, adding each rule to the model. The hydrology NER model outlined above is similar to NLTK's part-of-speech tagger and chunking modules (Yzaguirre, Smit, & Warren, 2016) and Amazon Comprehend (Zarei & Nik-Bakht, 2019). This model uses a convolutional neural network (CNN) with four different units. The first unit of the model extracts features from the annotated dataset by modeling every word with their normalized text, prefix, suffix, and shape, then inserting them into an embedding table. The second unit takes the features from the embedding table and applies a 1-dimensional convolution filter along the text sentence. The third unit summarizes and pools the convolution layer. Finally, the fourth unit is a standard multi-layer perceptron (MLP) that takes the summarized features and predicts the annotated entity (Strubell, Verga, Belanger, & McCallum, 2017). With these units, the spaCy model feeds the input data forward and initializes the weights, sends errors back with backward propagation, and adjusts the values with an optimizer (such as Adam: Kingma & Ba, 2014). To train the model, we run spaCy's Command Line Interface train command over our training and development datasets. This ensures the trained model balances maximizing the accuracy of the model, and overfitting to the training set. To minimize the impact of sample bias and improve the classification process, we also incorporate domain adaptation as part of the model training (Kouw and Loog, 2019). Here, we instantiate the custom spaCy model with the state of a generic pre-trained model (spaCy, 2021b) that is able to accurately extract tags such as 'PEOPLE', 'ORG', 'GEP', and 'PRODUCT'. Overall, by combining the results of the rule-based matching, a machine learning model, and domain adaptation, we achieve a hybrid model that is able to extract entities from the text and performs better than either of the techniques used independently (Section 5.1).

4.2. Relationship Extraction, Entity Linking, and Geocoding

In order to get the precise locations of events and flood risk reduction projects, we first perform RE between street and city entity pairs extracted from the NER model, then geocode the pairs. To do this, we examine two methods: a nearest neighbour and a nearest syntactic dependency heuristic. In nearest neighbour heuristic, we use the proximity of words to select the location that occurs nearest

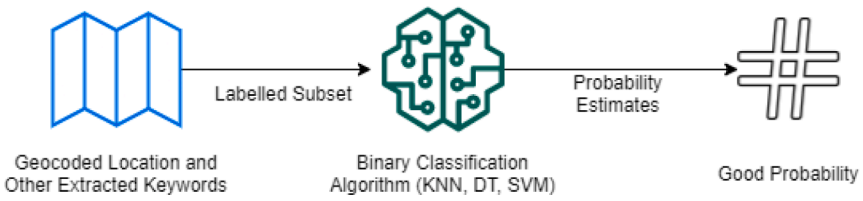

Fig. 4. High level Binary Classification Process.

to the flood reason entity in the text. We assume that the closer the proximity of the words are, the more likely they should be linked together. For example, given the sentence 'A major storm briefly closed down Warren Street in New York City, and both Grand and Washington Street in Jersey City', we can interpret that Warren Street is in New York City, while Grand Street and Washington Street are in Jersey City. However, if we were to find the closest city for Grand Street using word position distance, Grand is 4 words away from the New York City entity, and 7 words away from the Jersey City entity, causing Grand Street to be linked with New York City. Using syntactic dependency instead, New York City is 4 jumps away from Grand, while Jersey City is only 2 jumps away, correctly linking the two entities (see Fig. 2). As part of our validation process, we evaluate the effectiveness of the two RE heuristics in Section 5.2.

To implement the heuristic, we use Dijkstra's algorithm, or the single-source shortest path undirected graph algorithm. Dijkstra's algorithm allows us to find the shortest path between a target and source node in a graph, the same goal as the syntactic dependency heuristic. In our case, for each article we link each street entity to the nearest syntactic flood reason and city entity for the geocoding process. Specifically, using spaCy's dependency parser, we convert each sentence in an article or document into a dependency tree. We then convert the tree into a graph with each node as a word and link the sentence heads to allow cross-sentence extraction. Fig. 3 shows a sample paragraph and the application of the dependency parser linking all the sentences together at the head of the dependency tree. Finally, we convert the tree into a graph and apply Dijkstra's algorithm using NetworkX's (NetworkX — NetworkX 2021) single-source shortest path undirected graph function.

Prior to the geocoding process we include an EL step by removing city entities not in the pre-defined list of city/state pairs (Denis, 2014). For example, even if the model correctly identifies "Sioux Falls" as a city, but is linked with the state "Tennessee", the city is skipped. With the remaining linked street, city, and sub-region/state, we input the street/city pairs obtained from the Dijkstra's algorithm into geocodio (LLC, n.d.) and Google Maps API (Google, n.d.) to perform entity linking as well as geocoding. These services have been used by past researchers (Präger, Kurz, Böhm, Laxy, & Maier, 2019, Robinson & Steil, 2020, Kiaghadi, Rifai, & Dawson, 2021) and returns a lookup accuracy score between 0 and 1 which is used to further filter out poorly paired locations. Through geocoding, we apply EL and find the latitude / longitude coordinates for our extracted locations. The final evaluation of the RE and EL process can be found in Section 5.2.

For geocoding the NOAA Storm Event dataset, in addition to the above steps to link street, city, and sub-region/states, we also geocode major bodies of water by masking out the location that intersects with major water bodies. Since the NOAA Storm Event dataset only includes a county and sub-region/state of the events, we use the intersection of water bodies and county layer and take the mean coordinate to get the latitude and longitude.

4.3. Location Filtering

In the case of the historical floods and flood risk reduction project model, despite using different RE heuristics, city/state filtering, and geocoding EL, many of the linked and geocoded locations did not match the expected article location. To this end, a simple binary classification model seen in Fig. 4 was further trained using the manually labelled data, such as the number of steps from the flood entity to the city entity, to return a *'Good probability'* value. This value, between 0 and 1, with a threshold of 0.5, is used to rank the extracted locations from least confident to most confident. For the model selection, we compare the use of different machine learning algorithms such as decision trees, naive Bayes, and support vector machines in the scikit-learn library (scikit-learn, 2021a) to validate the geocoding process, and ultimately narrow the number of false positives. Specifically, using a manually labelled dataset of good and bad extractions from the NER model and entity linking (Section 4.4.2), for each algorithm we apply the Stratified K-Folds cross-validation from the scikit-learn library (scikit-learn, 2021b) with a split of 20, and measure the accuracy and learning curve scores. After filtering, we found the resulting dataset to be more accurate and less likely to show irrelevant articles and locations. The evaluation and final classification model details can be found in Section 5.3.

4.4. Data and Accuracy

4.4.1. Data Acquisition

Machine learning requires a large domain-specific dataset for the training of the model. To address this, and evaluate the accuracy of the NER and filter model, a dataset was collected and manually labelled. Firstly, use Meltwater (Meltwater: Media Monitoring & Social Listening Platform 2021)), a service that finds and provides an easy interface to filter and query news articles, to gather a dataset of publicly available articles. Specifically, we query the service for news articles in the United States between the years 2008 and 2019 with the keywords 'flooding', 'roads, 'hurricane', 'tidal', 'storm', 'rain', and 'river', using the foregoing filtration method. This process

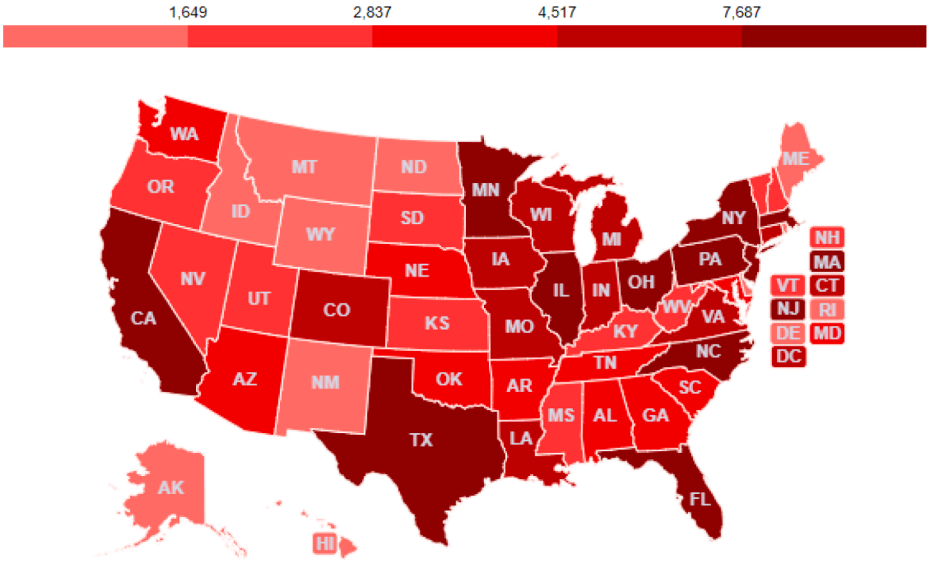

Fig. 5. Distribution of Newspaper Articles Sources Over the 54 States

Table 1
Augmentations and configurations used during training time of the NER model

Model Configuration	Description	Examples
Similar Replacement (SR)	The training set is duplicated and each entity of interest is replaced with another labeled	Original: "Texas was flooded" Augmented: "Texas was submerged"
City Injection	Include a list of tagged cities to the training set	Added to the training set: [{"orth": "Holtsville", "tag": "-", "ner": "U-cit"}, {"orth": "Wrangell", "tag": "-", "ner": "U-cit"}]
Training Sentence Filter	Filtering the training set to only sentences with at least one entity of interest	Sentence Removed: ["for immediate release contact: murray press office"]
Domain Adaptation	Use of a model trained with the generic NER task for the training of our NER model	Generic Model: "Tornadoes were also possible Saturday [DATE] morning over parts of southern Louisiana [GPE]"

initially yielded 653,409 article texts extracted, with news sources ranging from small county publishers such as the "Pike County Journal" (The Pike County Journal-Reporter, 2021) with short local road closures, to major newspaper publishers such as the "Omaha World-Herald" (Omaha.com, 2021) covering long pieces on events. Unfortunately, nearly half of the articles in the dataset resulted in a download error or no data was received from the article site, which significantly skews the accuracy and reliability of the article corpus to articles that do not block the downloading of the article content. These errors were removed from the process, resulting in a corpus size of 323,052 articles. However, not all articles from the website were relevant to historical floods or flood risk reduction projects. Examples include sentences such as "no matter how good your cease-fire is if you keep flooding the place with arms on all sides" (Nevins, 2015) or "a pipe ruptured inside the walls of the Olmstead, a sprawling new luxury building just off Franklin Avenue in Crown Heights." (Offenhartz, 2019). In both cases, while these articles have the word or variant of the word 'flood', they are not relevant in the task of extracting historical flood events or flood risk reduction projects. These irrelevant articles are filtered out during the NER process, by removing any articles with no flooding entities. This further removes 72,308 articles that do not mention floods, resulting in a final total of 250,744 articles that are directly related to flooding. Despite the shortcomings, the resulting corpus maintains a large variety of newspaper publishers, and Fig. 5 shows the distribution of the newspaper articles over the 51 states. This figure shows that the least number of articles were from Wyoming (483 articles), and the most articles from California (23,278 articles), generally following the population distribution.

The NOAA Storm dataset was gathered from the Storm Event Database (NOAA 2021) from 1996 to 2019. This dataset contains 139, 141 events, with information such as the narrative of the storm event, the county, sub-region/state, loss of life, property damage, and floods resulting from the storm. Information from the dataset varies depending on the year due to data reformatting and standardization but overall is complete for the dataset.

4.4.2. Training Set and Data Augmentation

The dataset used for the NER model consisted of 134,644 words that were manually labeled and tagged with 18 relevant entity labels using a random sample of 66 articles gathered from meltwater. The tags used in the dataset are: tid (tidal floods), rai (rain / pluvial), riv (river / fluvial), hur (hurricanes), sto (storms), str (streets / roads), bui (named buildings), wat (bodies of water), dat

Table 2
Model comparisons for hybrid, rule-only, trained-only model with domain adaptation and data augmentation applied; the numbers represent the max F-score.

Model	F1 Score	Recall	Precision
Hybrid	**72.13%**	74.80%	69.47%
Rule-Only	52.53%	60.61%	44.44%
Trained-Only	69.62%	69.78%	69.47%

(dates), ada (adaptation measure type), met (metrics / return periods), cit (cities), ent (organizational entity), loc (county locations), prj (adaptation project name), cst (costs), flo (floods), and O (none of the above). For each word, the corresponding BILUO Scheme is used to encode the entity classes (Ratinov & Roth, 2009). The BILUO Scheme consists of 'B-cit' for the beginning of a city entity, 'I-cit' for the intermediate text of a city entity, 'L-cit' for the last text of the city entity, and 'cit' for a singular city entity, or a 'O' for words that were not relevant to the task. The most common tag was the location tag, 'loc', with 2,983 labels, followed by city and streets 'cit' and 'str' with 947 and 737 labels respectively. The least common tag was the return period tag 'ret' with only 12 tags. Metric and project are the next least common tag, with only 52 and 70 tags. The vast majority of the dataset is the irrelevant tag, 'O', which makes up '125,527′ of the '134,644′ words. As a result, only about 6.77% of the corpus contains relevant entities. Finally, due to the relatively small number of sample articles and relevant tagged words used for the dataset we use a split of 70% and 30%. With 66 articles worth of labelled words, a split of 70/30 provides 46 articles in the training set, and 20 in the test set. While typically a split of 2/3 training and 1/3 testing is used, 60/40, 70/30, 80/20 or 90/10 are also commonly used (Raschka, 2020) depending on the size of the dataset. However, with a higher training to test set ratio like 90/10, the number of test articles would only be 7 (6.6 rounded), making the justification of a representative sample difficult. Ultimately, the choice of the split strives to maximize the number of training samples, while maintaining a test set accuracy that is meaningful Table 1.

Due to the sparseness of this dataset, we used data augmentation techniques such as substitution and addition to expand the number of training sentences in the training set only. Other common data augmentation techniques including flipping, rotating, resizing, and scaling are simple ways to augment an image; however, these conventional methods are largely ineffective for text due to the complex nature of text. Unlike an image, it is challenging to come up with generalized rules for language transformation (Wei & Zou, 2019). By using conventional methods, such as flipping, the entire sentence can be taken out of context or become grammatically incorrect, which ultimately defeats the purpose of building syntactic dependency into the model to improve accuracy. Since the goal of the NER model is to learn the patterns of the text and not memorize specific words, the data augmentation used in this task substitutes words and phrases in the sentences with other words that maintain similar meaning, while generating new data that retains meaningful sentence structure. For example, we perform a more generalized version (which we call similar replacement) of synonym replacement, where synonym replacement involves "Randomly choose n words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random" (Wei & Zou, 2019, Kolomiyets et al., 2011; Zhang et al., 2015; Wang and Yang, 2015). Using this more generalized technique, we augment a new sentence "Thousands of Texans have been impacted by the devastating floods brought on by Hurricane Irene." from the original sentence "Thousands of Texans have been impacted by the devastating floods brought on by Hurricane Harvey." While the impact on the accuracy due to data augmentation is minimal (see Section 5.1 Table 4), data augmentation is important in reducing overfitting by preventing the model from memorizing and tagging specific words, and instead learn the patterns and context for the sentence of interest. Furthermore, through data augmentation we introduce additional words not originally found in the training set into the model vocabulary, priming the model to pick out words that may exist in the test set but not in the original training set. Finally, we also introduce the tagged names of over 60,000 cities and city aliases in the United States (Denis, 2014) into the training data at least once to aid the word embedding and help the model identify cities that are not in the original training dataset. Aliases include "Nyc", and "Ny City" for New York City, and neighborhood names such as "Manhattanville" and "Lenox Hill". With this dataset, we have the necessary input to train a statistical NER model for entity extraction.

4.4.3. Evaluation Metric

In this analysis, we use the F1 score to evaluate the accuracy of the NER model, and overall process in extracting the correct entities from the text. The F1 score (Eq-1) takes the precision (Eq-2) and recall (Eq-3) accuracy to measure the accuracy of the model.

$$F1 = 2 \cdot \frac{(Precision * Recall)}{(Precision + Recall)} \tag{1}$$

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2}$$

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3}$$

Here, a low precision means that there are many more false positive entities, resulting in irrelevant words to be incorrectly tagged as an entity. A low recall on the other hand indicates that there are many more false negative entities, resulting in entities that should have been tagged to be skipped. Using these two measures, the F1 score (Eq. 1) strikes a balance between tagging the right entities, and

Table 3
Model comparisons for domain adaptation only; the numbers represent the max F-score across 30 iterations

		Domain Adaptation Only	
		No	Yes
Model	Historical	66.51%	**69.41%**
	Adaptation	64.04%	**68.22%**

Table 4
Model comparisons with only data augmentation and domain adaptation on the trained-only model; the numbers represent the max F-score across 30 iterations

		Similar Replacement			
		n=0	n=1	n=2	n=3
City Injection	**n=0**	69.414%	69.565%	68.481%	68.142%
	n=1	68.722%	68.889%	66.225%	**69.623%**
	n=2	67.55%	67.532%	69.451%	67.841%
	n=3	68.709	66.957	67.544	67.974

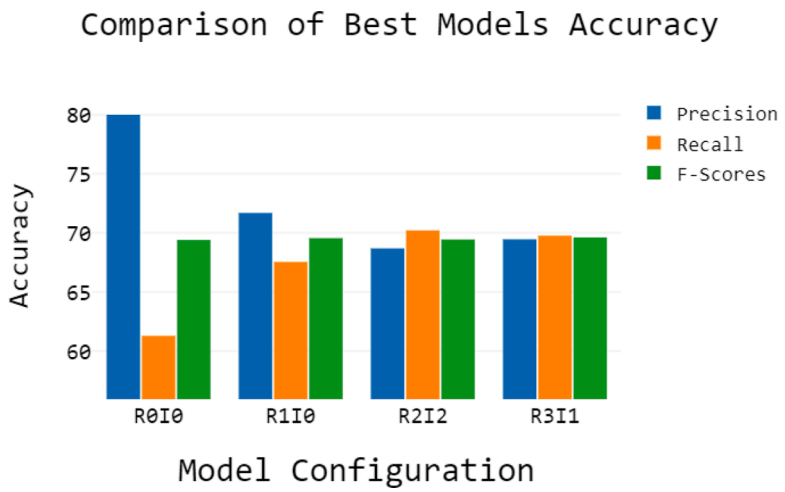

Fig. 6. F1, Recall, and Precision of NER Configurations for Historical Floods with only data augmentation and domain adaptation on the trained-only model

Table 5
Model comparisons for relationship extraction on the number of locations extracted

	Relationship Extraction	
	Closest Entity	Dependency Parser
Before Filter	195,484	181,076
After Filter	27,444	54,988

not missing any entities. The F1 score is then compared for different configurations of the NER model and data augmentations to identify the most accurate model. The different configurations for the model can be seen in Tables 2, 3, and 4 in the results section.

The fraction correct (FC) score or the accuracy score (scikit-learn, 2021b) is applied to evaluate the binary classification and compare different machine learning algorithms (such as SVM, decision tree, random forest). This measure was then complimented with the learning curves which allowed us to evaluate each algorithm's potential for improvement with increased data points. The learning curve is defined as a plot between the accuracy measure of the training and validation sets and the number of labeled samples. By observing the curve's slopes, we are able to deduce whether or not an algorithm can continue to improve, or if it has already converged with the given data. Overall, the combination of the learning curve and accuracy measure allows us to isolate the algorithm that performs the best with the given data.

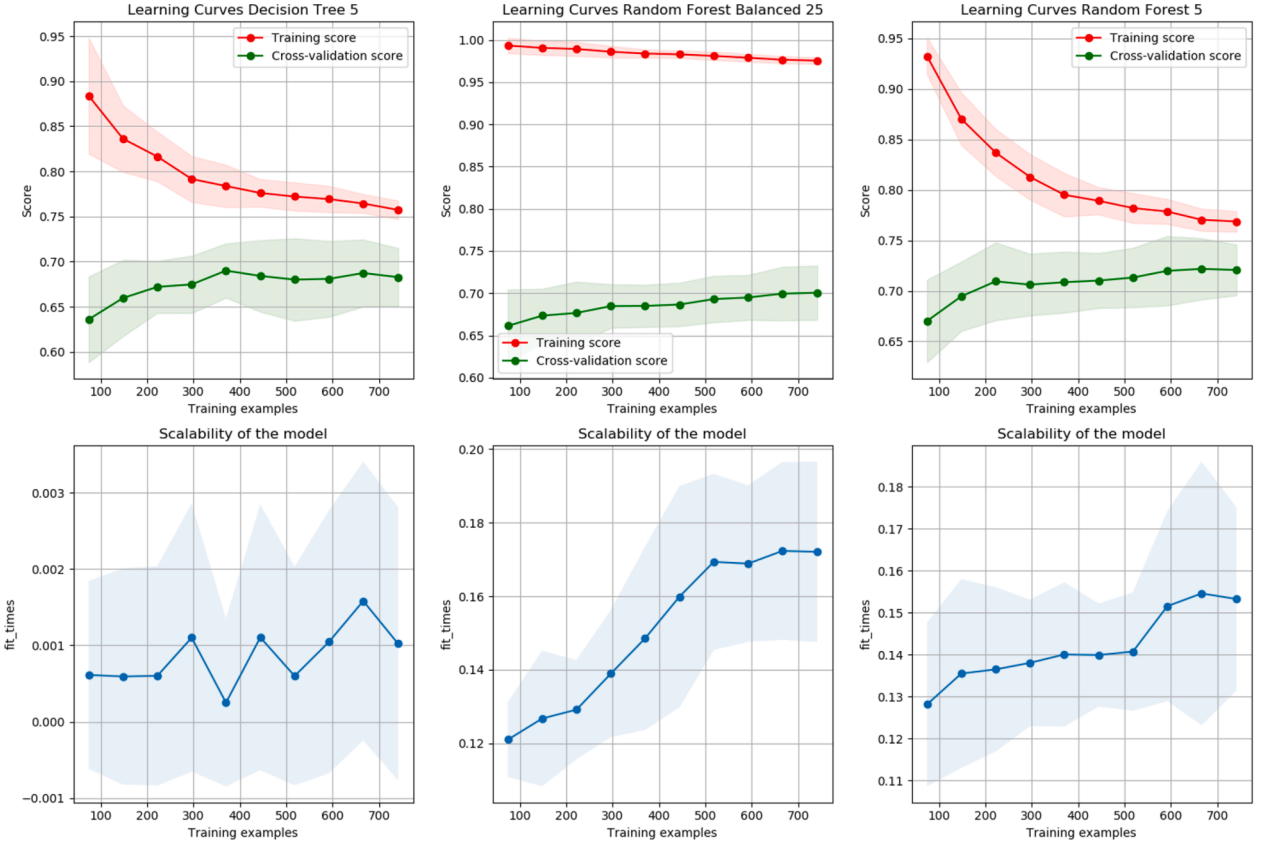

Fig. 7. Learning Curve and Scalability of Binary Classification Top 3 Filter Models

Table 6
The results of NER model extraction.

	Historical Floods Model	Flood Risk Reduction Project Model	NOAA Storm Model
Corpus Size	653,409 articles	653,409 articles	139,141 Storm Events
Number of Entities Extracted	6,203,728	5,689,440	762,948
Number of Locations Extracted	195,484	181,076	435,353
Number of Locations After Filtering	27,444 (14%)	54,988 (30.36%)	-

5. Results

In this section, we describe and compare the results of applying different configurations to the NER model, RE, EL, and location filtering with the dataset collected and labelled from Meltwater.

5.1. Evaluation of NER Model

Training a custom hybrid spaCy NER model, we achieved an F1 score of 0.7213 by applying domain adaptation, 3x similar replacement, and 1x city injection (see Table 2) to the test set. Comparing this result with just a rule-based model, we found an increase of 20.85% (rule-based had an F1 score of 0.5128), while compared to a trained machine learning model there was an increase of 2.51% (trained-only model had an F1 score of 0.6962).

Comparing the different configurations applied to our model, our results indicate that firstly applying a general domain adaptation (DA) can improve the model performance by 3% - 4% (See Table 3). Secondly, on top of DA, applying data augmentation by having a single sample in the city label class and three similar replacements adds a slight increase in performance (0.2%) than none sample or multiple sample (see Table 4 and Fig. 6), likely due to adding unseen cities and entities to the training set in the former and overfitting to the augmented data in the latter.

5.2. Evaluation of Relationship Extraction and Geocoding

To compare the effectiveness of using syntactic dependency and neighbors of words, we first examine the number of extracted locations. Applying both methods on all 653,409 articles, we find that running syntactic dependency using the dependency parser followed by geocodio extracts 181,076 locations, while neighbors of words extracts 195,484 locations. However, although neighbors of words extracts more locations, we also see that after applying location filtering only 27,444 (12.67%) of the locations remain, whereas 54,988 (30.36%) locations remain when using syntactic dependency (see Table 5). Using a set of 928 manually labelled articles, we conclude that despite extracting less entities, the locations found with syntactic dependency are more contextually accurate. For the geocoding process, the majority (94.66%) of the lookup accuracy values returned by the services were 1 or 0.9. A further analysis of 50 random locations found 35/50 (70%) of the locations returned by the geocoder within an acceptable range of the input location.

5.3. Evaluation of Location Filtering

Testing against the same 928 manually labelled articles as the RE evaluation, the results of our binary classification process yields a 73% accuracy in extracting the correct location from an article, using the Random Forest (RF) algorithm with a max depth of 5. Tuning the algorithm using grid search, we found the best performance using a depth of 5 and balanced input classes. Fig. 7 presents the results of validation for the classification process. In addition to model performance diagnosis, the scalability curves indicate times required by the models to train incrementally. Here, we tested the nine different algorithms mentioned in Section 4.3. As seen in the figure of the top three performing algorithms, the Random Forest algorithm with a max depth of 5 is able to achieve a better rating accuracy than the other algorithms on the (green) validation set for any sample size greater than 200. We also compare the results of a Random Forest with a max depth of 25, and see the (red) training set accuracy is near 1 showing a case of overfitting. Finally, it is worth noting that while the scalability graphs for Decision Tree show that it is not monotonically increasing, the difference in the time is minimal and relatively constant.

5.4. Discussion of Process Results

Running the full process on the 653,409 articles in the article dataset, the Historical Flood NER model resulted in 6,203,728 entities and 195,484 historical flood locations (see Table 6). After applying the binary classification to the historical flood location dataset, 27,444 (14%) returned a “Good Probability” greater than 0.5. Similarly, running the flood risk reduction project NER model on the 653,409 articles in the article dataset resulted in 5,689,440 entities and 181,076 flood risk reduction project locations. Of these locations, only 54,988 (30.3%) of the flood risk reduction project location dataset had a “Good Probability” greater than 0.5. Fig. 8 shows the distribution of the classification output when applied to the extracted historical flood and flood risk reduction projects, with the historical flood skewed to the right, while the flood risk reduction project distribution is more uniform with peaks at the 0.5 threshold and extremities. The skewness of the historical flood distribution can likely be explained by the use of neighbors of words relationship

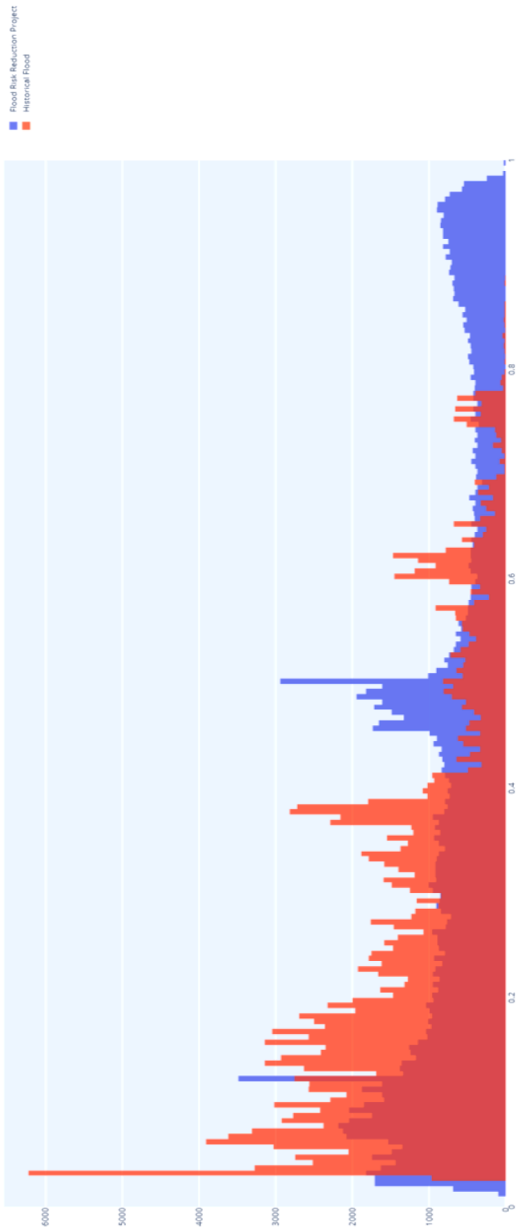

Fig. 8. Distribution of "Good Probability" from the Binary Classification

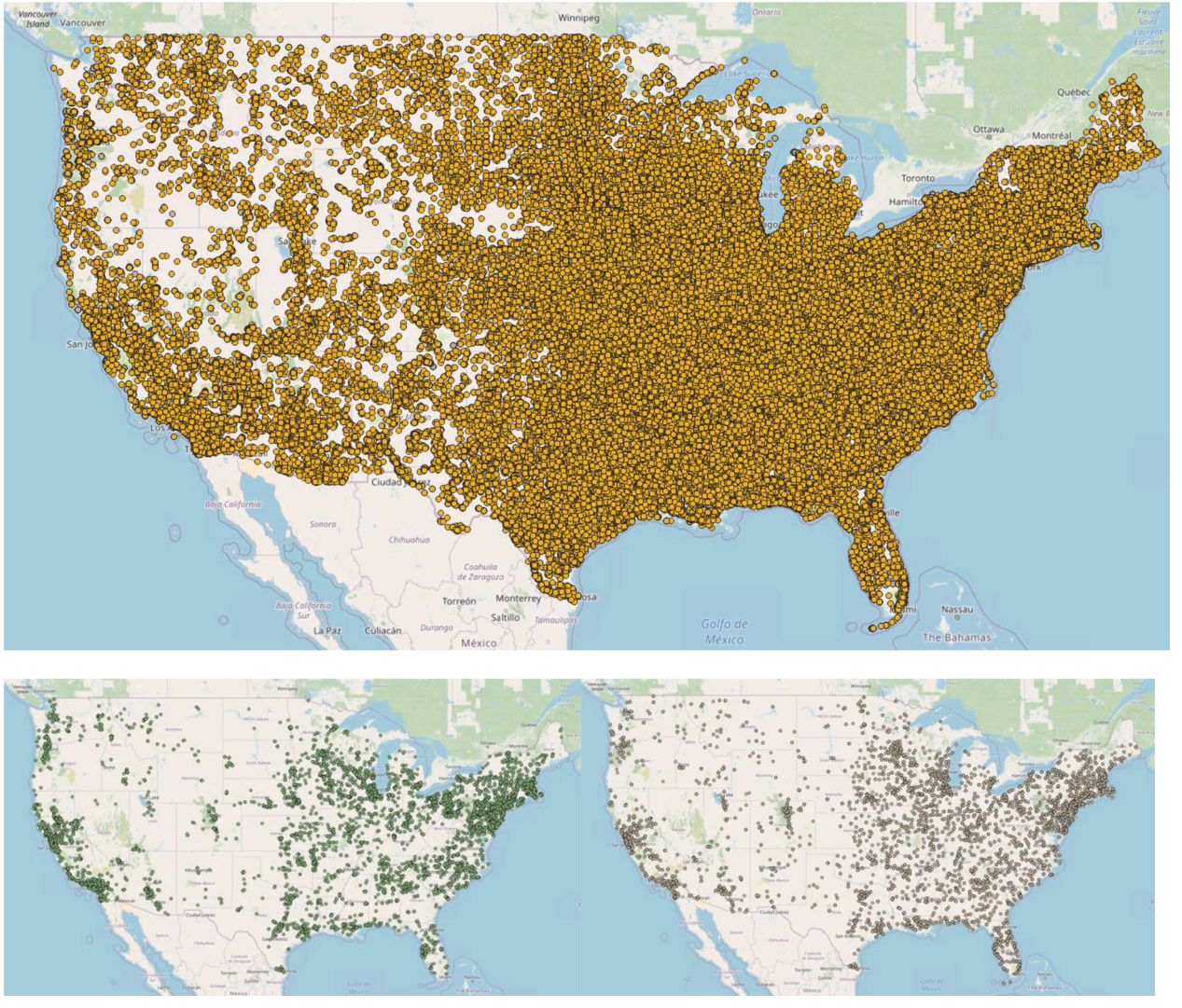

Fig. 9. Visualization of the Datasets. The top figure shows the storms extracted from the NOAA dataset. The bottom left figure shows the locations of historical floods, and the bottom right figure shows flood risk reduction projects

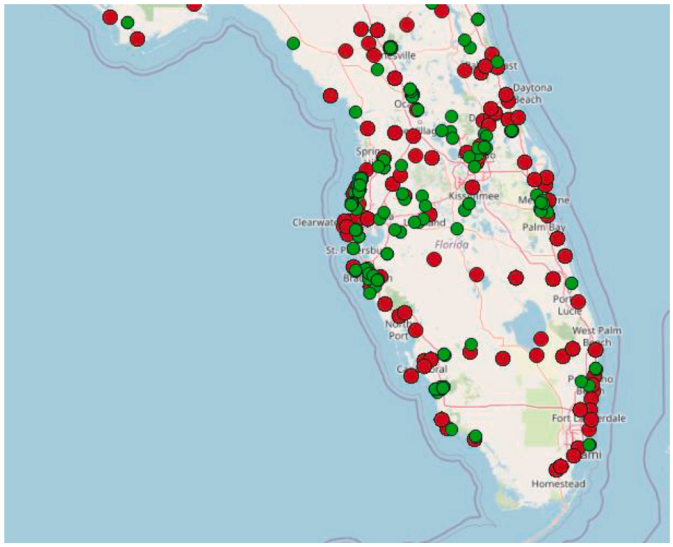

Fig. 10. A comparison of historical street floods (green) and existing flood risk reduction projects (red) in Florida (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

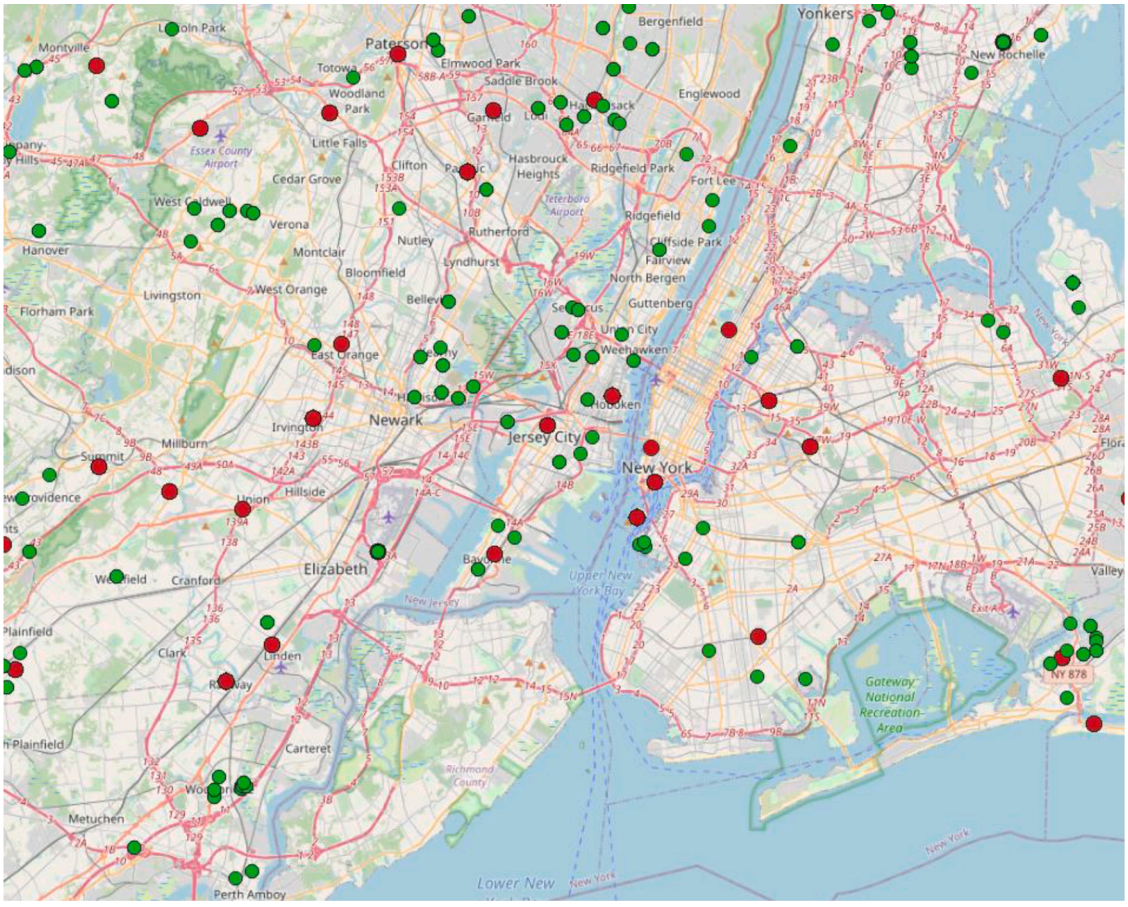

Fig. 11. A comparison of historical street floods (green) and existing flood risk reduction projects (red) in New York City (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

extraction, compared to the somewhat more uniform distribution of the flood risk reduction project resulting from the use of syntactic dependency.

The NOAA Storm NER Model was run on the 139,141 NOAA Storm Events. This resulted in 762,948 entities and 435,353 storm locations (see Table 6). Of the locations found, 326,724 locations, or 75.0%, had a valid geometry point. Of these valid points, 169,167 points, approximately 51.8%, were supplied directly from the CSV. 16,120 points, approximately 4.9%, were from getting the mean coordinate by intercepting the location with the county map. Lastly, the remaining 141,437 points, about 43.3%, were from geocoding using Google or geocodio.

When compared to human validation our model achieved an F1 score that was 8.4% lower than the human. To run this test, we collected 35 random articles with 183 locations from the corpus and applied the full NER model, geocoding, and binary classification to each article, extracting streets that have flooded and their locations. We then benchmarked the model against a data analyst, who evaluated the same articles and also extracted flooded streets. Finally, we compared the results of the model and the data analyst to a gold-standard expert human evaluator with access to external location lookup tools and knowledge outside the individual article. Our results found that while the model was only able to achieve an F1 score of 51.77%, the human data analyst scored an F1 score of 60.18% compared to the gold-standard. Specifically, we have found that the human identified 80% more false positives of flooded roads than the model, while the model had 31% more false negatives than the human, which is likely attributed to the binary classification model filtering out unlikely candidates.

We have identified four primary reasons as to why this process did not perform as well as expected. First, in many cases, the street, city, and sub-region did not match up to a real location. While the nearest syntactic dependency heuristic performed better than the nearest neighbour heuristic, the heuristic is not perfect and further research in correctly connecting streets to cities is needed to refine the process. Secondly, since the news article text comes from a variety of different news sources, the written format and patterns greatly differ from each other. While a trained model is expected to recognize good patterns that span across the corpus, more refined rules and training samples would help increase the accuracy and generalizability of the model. Thirdly, the binary model used to determine if a location is good or not itself is only 73% accurate, misclassifying many locations even in just that step. Finally, as seen in the results of the human validator scoring only 60.18% against the gold-standard with many false negatives, it is extremely difficult in many articles to accurately identify street, city, and flood reasons without external information outside the original article. A further look into entity disambiguation to verify extracted locations may prove to be beneficial in improving the accuracy.

5.5. *Implications*

The implications of this dataset are twofold. Firstly, the dataset provides new data for climate researchers to identify and map flood patterns, and to evaluate existing flood models. Since the dataset is geographically wide, it can be used to ensure flood models are unbiased, are not overfitting, and accurately represent flood risks. Secondly, Fig. 9's historical NOAA storm data map can be used to determine flood risk for land development and future storm water drainage. Another example of future work is the comparison of the data between historical floods (Fig. 9, bottom left) and existing flood reduction projects (Fig. 9, bottom right) to inform decisions on where to best build future flood risk reduction projects for maximum impact. Further insight can be seen in Figs. 9, 10, and 11, where Florida Panhandle has many flood risk reduction projects, while New York City has substantially less defenses, possibly making the city more susceptible to future flooding. Finally, future research using this data may indicate the biased flood effects on different communities and marginalized groups by comparing areas with high historical floods and no flood risk reduction projects to areas that are high risk but protected.

6. Conclusion

Although gathering nation-wide flood data is difficult, in this project we have been able to accomplish the creation of a detailed nation-wide historical flood, flood risk reduction project, and storm database using recent techniques in information extraction. From the extraction of entities using an NER model, to the entity extraction and geocoding, and finally the filtering of locations, we achieved an F-score of 72.13% on the entity extraction, an accuracy score of 73% on the binary classification filter, and an overall process F-score only 8.4% lower than a human validator. Through this process, we have extracted a total of 27,444 streets with flood instances, 54,988 flood risk reduction projects, and 435,353 storms over the nation in the past decade, with additional auxiliary details such as the flood cause and project costs. We also perform a comparison between the nearest neighbor heuristic and the syntactic neighbor heuristic, and identify 27,544 additional entities using the syntactic neighbor heuristic. That said, the project can be improved and expanded upon. For example, it may be possible to use both social media and newspaper articles to boost the number of reported flood events, while cross-reference flood events and extracting detailed flooding information from newspapers. In terms of the methodology used in this paper, the hybrid model could be further improved by reducing the number of false positives and duplicate locations. Furthermore, the use of the nearest syntactic dependency heuristic should be re-examined, and possibly replaced with a more accurate heuristic. Following this improvement, the binary classification model used to filter out incorrect locations could also be improved from the current 73% achieved by the simple Random Forest algorithm. Despite these improvements, the data gathered through this project may prove to be invaluable, and all-in-all through this project we have created an alternative database to government sources and commercial databases that can be used for future flood pattern modeling, and analysis.

References

Ahmad, K., Pogorelov, K., Riegler, M., Ostroukhova, O., Halvorsen, P., Conci, N., & Dahyot, R. (2019). Automatic detection of passable roads after floods in remote sensed and social media data. *Signal Processing: Image Communication, 74*, 110–118. https://doi.org/10.1016/j.image.2019.02.002

Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. P. (2018). Social sensing of floods in the UK. *PLOS ONE, 13*(1), Article e0189327. https://doi.org/10.1371/journal.pone.0189327

ASCE. (2017). ASCE's 2017 Infrastructure Report Card. Retrieved from https://www.infrastructurereportcard.org/cat-item/levees/.

ASFPM _MaptheNation_Report_2020.pdf. (2020). Retrieved from https://asfpm-library.s3-us-west-2.amazonaws.com/FSC/MapNation/ASFPM_MaptheNation_Report_2020.pdf.

Bischke, B., Bhardwaj, P., Gautam, A., Helber, P., Borth, D., & Dengel, A. (2017). Detection of Flooding Events in Social Multimedia and Satellite Imagery using Deep Neural Networks. 3. Retrieved from http://ceur-ws.org/Vol-1984/Mediaeval_2017_paper_51.pdf.

de Bruijn, J. A., de Moel, H., Jongman, B., de Ruiter, M. C., Wagemaker, J., & Aerts, J. C. J. H. (2019). A global database of historic and real-time flood events based on social media. *Scientific Data, 6*(1), 311. https://doi.org/10.1038/s41597-019-0326-9

Denis. (2014). Grammakov/USA-cities-and-states. Retrieved from https://github.com/grammakov/USA-cities-and-states (Original work published 2014).

Dias, M., Boné, J., Ferreira, J. C., Ribeiro, R., & Maia, R. (2020). Named Entity Recognition for Sensitive Data Discovery in Portuguese. *Applied Sciences, 10*(7), 2303. https://doi.org/10.3390/app10072303

Dutt, R., Basu, M., Ghosh, K., & Ghosh, S. (2019). Utilizing microblogs for assisting post-disaster relief operations via matching resource needs and availabilities. *Information Processing & Management, 56*(5), 1680–1697. https://doi.org/10.1016/j.ipm.2019.05.010

Gabbard, R., DeYoung, J., Lignos, C., Freedman, M., & Weischedel, R. (2018). Combining rule-based and statistical mechanisms for low-resource named entity recognition. *Machine Translation, 32*(1), 31–43. https://doi.org/10.1007/s10590-017-9208-0

Ghafarian, S. H., & Yazdi, H. S. (2020). Identifying crisis-related informative tweets using learning on distributions. *Information Processing & Management, 57*(2), Article 102145. https://doi.org/10.1016/j.ipm.2019.102145

Gudivada, V. N., & Arbabifard, K (2018). Chapter 3—Open-Source Libraries, Application Frameworks, and Workflow Systems for NLP. In V. N. Gudivada, & C. R. Rao (Eds.), *Handbook of Statistics* (pp. 31–50). Elsevier. https://doi.org/10.1016/bs.host.2018.07.007.

Ji, D., Tao, P., Fei, H., & Ren, Y. (2020). An end-to-end joint model for evidence information extraction from court record document. *Information Processing & Management, 57*(6), Article 102305. https://doi.org/10.1016/j.ipm.2020.102305

K, S., & Thilagam, P. S (2019). Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers. *Information Processing & Management, 56*(6), Article 102059. https://doi.org/10.1016/j.ipm.2019.102059

Karimzadeh, M., Pezanowski, S., MacEachren, A., & Wallgrün, J. O. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation: GeoTxt: A scalable geoparsing system. *Transactions in GIS, 23*. https://doi.org/10.1111/tgis.12510

Kaufhold, M.-A., Bayer, M., & Reuter, C. (2020). Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management, 57*(1), Article 102132. https://doi.org/10.1016/j.ipm.2019.102132

Kiaghadi, A., Rifai, H. S., & Dawson, C. N. (2021). The presence of Superfund sites as a determinant of life expectancy in the United States. *Nature Communications, 12*(1), 1947. https://doi.org/10.1038/s41467-021-22249-2

Kolomiyets, O., Bethard, S., & Moens, M.-F. (2011). Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, 271–276*. Association for Computational Linguistics.

Kouw, W. M., & Loog, M. (2019). An introduction to domain adaptation and transfer learning. *ArXiv:1812.11806 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1812.11806.

Kozlowski, D., Lannelongue, E., Saudemont, F., Benamara, F., Mari, A., Moriceau, V., & Boumadane, A. (2020). A three-level classification of French tweets in ecological crises. *Information Processing & Management, 57*(5), Article 102284. https://doi.org/10.1016/j.ipm.2020.102284

Kron, W., Steuer, M., Löw, P., & Wirtz, A. (2012). How to deal properly with a natural catastrophe database – analysis of flood losses. *Natural Hazards and Earth System Sciences, 12*(3), 535–550. https://doi.org/10.5194/nhess-12-535-2012

Laumer, D., Lang, N., van Doorn, N., Mac Aodha, O., Perona, P., & Wegner, J. D. (2020). Geocoding of trees from street addresses and street-level images. *ISPRS Journal of Photogrammetry and Remote Sensing, 162*, 125–136. https://doi.org/10.1016/j.isprsjprs.2020.02.001

Liu, X., Kar, B., Montiel Ishino, F. A., Zhang, C., & Williams, F. (2020). Assessing the Reliability of Relevant Tweets and Validation Using Manual and Automatic Approaches for Flood Risk Communication. *ISPRS International Journal of Geo-Information, 9*(9), 532. https://doi.org/10.3390/ijgi9090532

Meltwater: Media Monitoring & Social Listening Platform. (2021). Retrieved from Meltwater website: https://www.meltwater.com/en.

Moore, F. C., & Obradovich, N. (2020). Using remarkability to define coastal flooding thresholds. *Nature Communications, 11*(1), 530. https://doi.org/10.1038/s41467-019-13935-3

Morss, R. E. (2010). Interactions among Flood Predictions, Decisions, and Outcomes: Synthesis of Three Cases. *Natural Hazards Review, 11*(3), 83–96. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000011

Munnelly, G., & Lawless, S. (2018). Constructing a knowledge base for entity linking on Irish cultural heritage collections. *Procedia Computer Science, 137*, 199–210. https://doi.org/10.1016/j.procs.2018.09.019

NetworkX — NetworkX Network Analysis in Python (2021). Retrieved from https://networkx.org/.

Nevins, S. (2015, February 2). How The US, Its Allies And Syria Unwittingly Corporatized ISIS. Retrieved from MintPress News website: https://www.mintpressnews.com/how-the-us-its-allies-and-syria-unwittingly-corporatized-isis/201748/.

NLP-progress. (2021a). Entity Linking. Retrieved from NLP-progress website: http://nlpprogress.com/english/entity_linking.html.

NLP-progress. (2021b). Relationship Extraction. Retrieved from NLP-progress website: http://nlpprogress.com/english/relationship_extraction.html.

NOAA. (2021). Storm Events Database. 2021 Retrieved from https://www.ncdc.noaa.gov/stormevents/.

Offenhartz, J. (2019, November 15). 'They Don't Care About Us': Low-Income Tenant In Luxury Building Says She's Been Left In The Cold. Retrieved from Gothamist website: https://gothamist.com/news/crown_heights_luxury_building_no_heat.

Oh, O., Kwon, K. H., & Rao, H. R. (2010). An exploration of social media in extreme events: Rumor theory and twitter during the HAITI earthquake 2010. In *ICIS 2010 Proceedings - Thirty First International Conference on Information Systems. Presented at the 31st International Conference on Information Systems, ICIS 2010. Retrieved from.* https://asu.pure.elsevier.com/en/publications/an-exploration-of-social-media-in-extreme-events-rumor-theory-and.

Omaha.com. (2021). 2021 Omaha News. Retrieved from Omaha.com website: https://omaha.com/news/.

Oral, B., Emekligil, E., Arslan, S., & Eryiğit, G. (2020). Information Extraction from Text Intensive and Visually Rich Banking Documents. *Information Processing & Management, 57*(6), Article 102361. https://doi.org/10.1016/j.ipm.2020.102361

Ortiz-Martínez, Y., & Jiménez-Arcia, L. F. (2017). Yellow fever outbreaks and Twitter: Rumors and misinformation. *American Journal of Infection Control, 45*(7), 816–817. https://doi.org/10.1016/j.ajic.2017.02.027

Präger, M., Kurz, C., Böhm, J., Laxy, M., & Maier, W. (2019). Using data from online geocoding services for the assessment of environmental obesogenic factors: A feasibility study. *International Journal of Health Geographics, 18*(1), 13. https://doi.org/10.1186/s12942-019-0177-9

Raschka, S. (2020). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *ArXiv:1811.12808 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1811.12808.

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *CoNLL 2009 - Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (pp. 147–155). Association for Computational Linguistics (ACL). https://doi.org/10.3115/1596374.1596399.

Robinson, D., & Steil, J. (2020). Eviction Dynamics in Market-Rate Multifamily Rental Housing. *Housing Policy Debate, 0*(0), 1–23. https://doi.org/10.1080/10511482.2020.1839936

Rossi, C., Acerbo, F. S., Ylinen, K., Juga, I., Nurmi, P., Bosca, A., …, & Alikadic, A. (2018). Early detection and information extraction for weather-induced floods using social media streams. *International Journal of Disaster Risk Reduction, 30*, 145–157. https://doi.org/10.1016/j.ijdrr.2018.03.002

Ruokolainen, T., Kauppinen, P., Silfverberg, M., & Lindén, K. (2020). A Finnish news corpus for named entity recognition. *Language Resources and Evaluation, 54*(1), 247–272. https://doi.org/10.1007/s10579-019-09471-7

scikit-learn. (2021a). Scikit-learn. Retrieved from https://scikit-learn.org/stable/modules/classes.html.

scikit-learn. (2021b). Scikit-learn. Retrieved from https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score.

Smith, A. (2020). 2010–2019: A landmark decade of US. billion-dollar weather and climate disasters. *National Oceanic and Atmospheric Administration*. Retrieved from https://www.climate.gov/news-features/blogs/beyond-data/2010-2019-landmark-decade-us-billion-dollar-weather-and-climate.

Smith, L., Liang, Q., James, P., & Lin, W. (2017). Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management, 10*(3), 370–380. https://doi.org/10.1111/jfr3.12154

spacy. (2021a). 2021 SpaCy • Industrial-strength Natural Language Processing in Python. Retrieved from https://spacy.io/.

spacy. (2021b). 2021 SpaCy English starters. Retrieved from English website: https://spacy.io/models/en-starters.

Strubell, E., Verga, P., Belanger, D., & McCallum, A. (2017). Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2670–2680). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1283.

Technical Mapping Advisory Council (TMAC) Annual Report 2015. (2015). *Annual Report*, 177. Retrieved from https://www.fema.gov/sites/default/files/documents/fema_tmac_2015_annual_report.pdf.

The Pike County Journal-Reporter. (2021). Retrieved from http://www.pikecountygeorgia.com/.

US Department of Commerce, N. (2019a). Weather Related Fatality and Injury Statistics. Retrieved from https://www.weather.gov/hazstat/.

US Department of Commerce, N. (2019b). Retrieved from https://www.weather.gov/media/hazstat/80years.pdf.

Wang, R.-Q., Mao, H., Wang, Y., Rae, C., & Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers & Geosciences, 111*, 139–147. https://doi.org/10.1016/j.cageo.2017.11.008

Wang, W. Y., & Yang, D. (2015, September). *That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets*. 2557–2563. https://doi.org/10.18653/v1/D15-1306.

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *ArXiv:1901.11196 [Cs]*. Retrieved from http://arxiv.org/abs/1901.11196.

Yzaguirre, A., Smit, M., & Warren, R. (2016). Newspaper archives + text mining = rich sources of historical geo-spatial data. *IOP Conference Series: Earth and Environmental Science, 34*, Article 012043. https://doi.org/10.1088/1755-1315/34/1/012043

Zarei, F., & Nik-Bakht, M. (2019). Automated Detection of Urban Flooding from News. *ISARC Proceedings*, 515–520. https://doi.org/10.22260/ISARC2019/0069

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (pp. 649–657). MIT Press.