A multi-modal approach towards mining social media data during natural disasters - A case study of Hurricane Irma

By: Somya D. Mohanty, Brown Biggers, Saed Sayedahmed, Nastaran Pourebrahim, Evan B. Goldstein, Rick Bunch, Guangqing Chi, Fereidoon Sadri, Tom P. McCoy, Arthur Cosby

Mohanty, S.D., Biggers, B., Sayedahmed, S., Pourebrahim, N., Goldstein, E.B., Bunch, R., Chi, G., Fereidoon, S., McCoy, T.P., & Cosby, A. (2021). A multi-modal approach towards mining social media data during natural disasters- a case study of Hurricane Irma. *International Journal of Disaster Risk Reduction*, *54*, 1-14. https://doi.org/10.1016/j.ijdrr.2020.102032

Abstract:

Streaming social media provides a real-time glimpse of extreme weather impacts. However, the volume of streaming data makes mining information a challenge for emergency managers, policy makers, and disciplinary scientists. Here we explore the effectiveness of data learned approaches to mine and filter information from streaming social media data from Hurricane Irma's landfall in Florida, USA. We use 54,383 Twitter messages (out of 784 K geolocated messages) from 16,598 users from Sept. 10–12, 2017 to develop 4 independent models to filter data for relevance: 1) a geospatial model based on forcing conditions at the place and time of each tweet, 2) an image classification model for tweets that include images, 3) a user model to predict the reliability of the tweeter, and 4) a text model to determine if the text is related to Hurricane Irma. All four models are independently tested, and can be combined to quickly filter and visualize tweets based on user-defined thresholds for each submodel. We envision that this type of filtering and visualization routine can be useful as a base model for data capture from noisy sources such as Twitter. The data can then be subsequently used by policy makers, environmental managers, emergency managers, and domain scientists interested in finding tweets with specific attributes to use during different stages of the disaster (e.g., preparedness, response, and recovery), or for detailed research.

Keywords: Data mining | Social media | Natural disaster | Machine learning

Article:

1. Introduction

Climate change is expected to drive increases in the intensity of tropical cyclones [1] and increase the occurrence of ‘blue sky’ flooding [2]. Despite these hazards, coastal populations [3], and investments in the coastal built environment [4] are likely to grow. Understanding the impact of extreme storms and climate change on coastal communities requires pervasive environmental sensing. Beyond the collection of environmental data streams such as river gages, wave buoys, and tidal stations, internet connected devices such as mobile phones allow for the creation of real-time crowd-sourced information during extreme events. A key area of research is understanding how to use streaming social media information during extreme events — to detect disasters, provide situation awareness, understand the range of impacts, and guide disaster relief and rescue efforts [e.g. 5–11].

Twitter – with approximately 600 million tweet posts every day [12] and programmatic access to messages – has become one of the most popular social media platforms, and a common data source for research on extreme events [e.g. 13–15]. In addition to text, a subset of messages shared across Twitter contain images captured by its users (20–25% of messages contain images/videos [16]). A key hurdle for studying these aspects of extreme events with Twitter is the data are both large and considerably noisier than curated sources such as dedicated streams of information (e.g., dedicated environmental sensors). Posts on Twitter during disasters might also be irrelevant, or provide mis- or dis-information [e.g.,17,18], highlighting the importance of filtering and subsetting social media data when used during disaster events. Therefore a key step in all work with Twitter data is to filter and subset the data stream.

Previous work has addressed filtering and subsetting Twitter data during hazards and other extreme events. Techniques have included relying on specific hashtags [e.g.,19], semantic filtering [20], keyword-based filtering [18], as well as natural language processing (NLP) and text classification that use machine learning algorithms [18, 21]. Classifiers such as support vector machine and Naive Bayes classifiers have been used to differentiate between real-world event messages and non-event messages [22], and to extract valuable "information nuggets" from Twitter text messages [23]. The tweets' length and textual features can also used to filter emergency-related tweets [23]. Tweets have been scored against classes of event-specific words (term-classes) to aid in filtering [18]. Previous work have filtered and subset tweets using expert-defined features of the tweet [24]. Images have also been used to subset tweets based on the presence/absence of visible damage [25]. Filtering can also be understood by the extensive work on determining the relevance of tweets for a given event — see recent work and reviews [26–28]. In the context of this paper, we view filtering as any generic process that subsets tweets, even beyond the binary class division of relevance.

A few studies have identified the significance of adding spatial features and external sources for a better assessment of tweets' relevance for disaster events. For example [29], enriched their model with geographic data to identify relevant information. Previous work has used spatio-temporal data to determine tweet relevance [21], or linked geolocated tweets to other environmental data streams [e.g.,30,31].

As observed from prior work, capturing situational awareness information from social media data involves a hierarchical filtering approach [32]. Specifically, researchers/interested stakeholders filter down the data from the noisy social media data stream to fit their specific use cases (such as, type of image - destruction, damage, flooding; type of text - damage, donation, resource request/offer; spread of information, etc). A key component in such an approach is the quality of baseline data capture. Towards this our study proposes a novel approach towards quality gating the data capture from the social media data streams using developed threshold measures. This baseline filtering methodology that can be used to find relevant tweets and refining the data capture routines. Specifically, the goal of our study is to explore a multi-modal filtering approach which can be used to provide situational awareness from social media data during disaster events. We develop an initial prototype using tweets from Florida, USA during Hurricane Irma. The filtering routine allows users to adjust four separate models to filter Twitter messages: a geospatial model, an image model, a user model, and a text model. All four models are tested separately, and can be operated independently or in tandem. This is a design feature as we envision the sorting and filtering thresholds will be different for different users, for different events, and for different locations. We work through each model and discuss the combined model in the following sections.

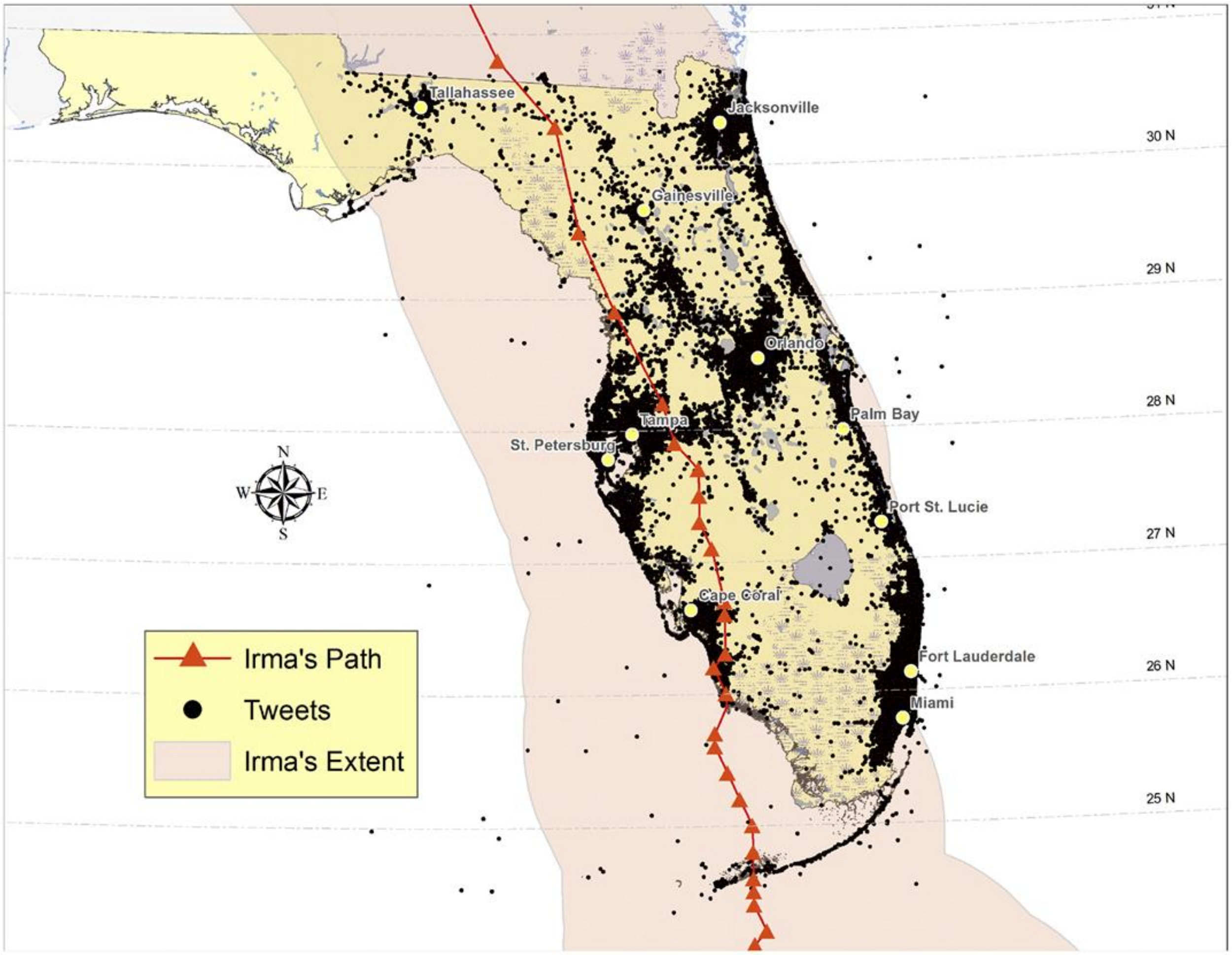


Fig. 1. The path of Hurricane Irma in September 2017 (orange line), the extent of tropical storm force winds (pink outline), and the location for all 784 K geolocated tweets used as the basis for this study (black dots).

2. Methodology

2.1. Hurricane Irma

Hurricane Irma (Fig. 1) was the first category 5 hurricane in the 2017 Atlantic hurricane season [33]. Hurricane Irma formed on August 31st, 2017, impacting many islands of the Caribbean, and finally dissipating over the continental United States [33]. Here we focus on the Twitter record of Irma specifically in Florida, USA. Irma made landfall in Florida Keys on October 09, 2017 as a category 4 hurricane and dissipated shortly after 09/13/2017. 134 fatalities were recorded as a result of the hurricane, with an estimated loss of $64.76 billion [34], making it one of the costliest hurricanes in the history of the United States [35].

2.2. Data collection and preprocessing

2.2.1. Twitter data

We used the Twitter Application Programming Interface (API) to collect tweets located in the geospatial bounding box that captured the state boundary of Florida. Tweets were recorded for the period of January 09, 2017 to October 10, 2017, and resulted in the collection of 784 K tweets from 96 K users during the time period. Our work is focused on 72 h (October 09, 2017 - December 09, 2017) when Hurricane Irma was near or over Florida. Therefore we subset the data and use 54,383 tweets from 16,598 users during this 72hr window. Fig. 1 highlights the locations of the Twitter messages, along with the path of Hurricane Irma, and the extent of tropical Storm force winds.

Each tweet from the Twitter API has 31 distinct metadata attributes [36] that can conceptually be grouped into three categories: 1) Spatio-temporal (time of creation and geolocation [latitude, longitude]), 2) Tweet content (tweet text, weblinks, hashtags, and images), and 3) Tweet source (account age, friends count, followers count, statuses count, and if verified). Geolocated tweets can have one of two types of location data — Places or Coordinates. Coordinates are exact locations with latitude and longitude attributes, while Places are locations within a Bounding Box or a Polygon designating a geospatial area in which the tweet is recorded [37]. For tweets with Places attributes, we transform the area representation to a single point by selecting the centroid of the Polygon as the location represented by the tweet. Within our study 42.58% (23,157) of the tweets had Coordinate locations and 57.42% (31,226) had Place locations.

2.2.2. Geospatial data

We collected meteorological sensor data, wind speed (in mph) and precipitation (in inches), for each county in Florida for the 72 h (October 09, 2017 - December 09, 2017). The hourly wind speeds was collected from the NOAA National Centers for Environmental Information (NCEI). Hourly precipitation values were obtained from the United States Geological Survey’s Geo Data Portal (USGS GDP) of the United States Stage IV Quantitative Precipitation Archive. Precipitation values from the closest weather station were used due to difficulty in obtaining reliable data for all weather stations. In addition to meteorological forcing, we collected data consisted of location of the hurricane’s eye, category of the hurricane, pressure and wind speed (NOAA National Hurricane Center). This data were discretized into hourly windows for the 72 h.

2.2.3. Data pre-processing

We aligned the 72hrs of Twitter data and the corresponding 72hrs of meteorological forcing data. Wind and precipitation values at the geolocation of a tweet was calculated using Inverse Distance Weighting (IDW). IDW is an interpolation method that calculates a value at a certain point using values of other known points:

$$W_p = \frac{\sum_{i=1}^{n} \frac{W_i}{D_i^k}}{\sum_{i=1}^{n} \frac{1}{D_i^k}} \tag{1}$$

where, W_p is the wind speed to be interpolated at point *p*, W_i is the wind speed at point *i*, D_i is the distance between point *p* and *i*, and *k* is a power function that reduces the effect of distant points as it goes up. IDW has been widely used to interpolate climatic data [38]. The method has demonstrated accurate results when compared to other interpolation methods especially in regions characterized by strong local variations [39]. IDW, for example, assumes that any measurement taken at a fixed location (e.g., weather station) has local influence on surrounding area, and the influence decreases with increasing distance. Within our study we chose IDW as our interpolation method as meteorological factors in a hurricane are highly influenced by local variations.

Furthermore, each tweet was also annotated with the corresponding temporal hurricane conditions data. Specifically, for each hourly time window, a tweet was associated with its distance from the eye of the hurricane and its conditions (i.e., pressure, max wind speed) during that window.

2.3. Multimodal scoring of tweet relevance

Our goal is to develop a single model for tweet relevance based on four sub models — 1) the relevance of the tweet based on geospatial attributes (i.e., the Tweets location relative to the forcing conditions of the hurricane, 2) the relevance of tweet images (when media is included in the tweet), 3) a score for the reliability of the user (i.e., network attributes to predict if a user is ‘verified’ by Twitter), and 4) the relevance of the tweet text. The methods used to construct of each of these models, and the results of models (submodels and the combined model) in are discussed in Section 3.

2.3.1. Geospatial model

Our goal in designing a geospatial relevance model was to search for thresholds in forcing conditions where tweets were likely to be related to Hurricane Irma, as opposed to background social media discussions. Specifically, we posit that the messages which are in close geospatial and temporal proximity to the disaster event will have more relevant situational awareness information than those which are not. Furthermore, such an approach can be used in real-time during the occurrence of an event where meteorological data can provide key information about disaster’s impact at different locations.

There are many meteorological conditions that can be used as proxies for extreme disaster conditions. We focus here on searching for modeling functions relating wind speed (w), precipitation (p), and distance from hurricane eye (d). We acknowledge that other factors could be used in addition to these three attributes. For example, rainfall during a given interval could be quantified in several ways, such as mean rainfall rate, max rainfall rate, total rainfall in a given interval. Similarly Wind metrics could include mean wind speed, max wind speed, metrics based on wind gusts, etc. For locations nearby the coast, metrics could include tide elevations, or storm surge elevations, and locations near streams could include stage and discharge data. Ultimately we chose Wind speed, precipitation and distance from the hurricane eye as these factors are available everywhere (vs metrics that are only applicable along streams and rivers) and because they are commonly available and collected by even basic meteorological stations. Nine different functions, - 1) $\frac{wind*rain}{distance}$, 2) $\frac{rain}{distance}$, 3) $\frac{wind}{distance}$, 4) $\frac{wind*rain}{\sqrt{distance}}$, 5) $\frac{rain}{\sqrt{distance}}$, 6) $\frac{wind}{\sqrt{distance}}$, 7) $\frac{wind*rain}{\sqrt[3]{distance}}$, 8) $\frac{rain}{\sqrt[3]{distance}}$, and 9) $\frac{wind}{\sqrt[3]{distance}}$, combining the geospatial attributes were compared to identify the best suited model towards creating a relevance score for the tweets. In each of the models, wind speed and precipitation acted as numerators (individually or combined), where as distance was used as a denominator — this was a heuristic method, as tweets are likely more relevant if forcing conditions are more severe (Higher wind, more precipitation, closer distance to hurricane)

Approximately 19,000 tweets from the Irma dataset were hand labeled by human coders as "Irma related" or "non-Irma related" based on the tweet content. The performance of each geospatial function was evaluated by comparing the ratio of Irma related tweets to total number of tweets during each time window. The ratios obtained from each formula was normalized using three different approaches - Min-max scaling, Log ($\log_{10}()$), and Box-Cox transformations. Ranking of Shapiro-Wilk (SW) test statistics was used to assess normality. In addition, multiple observed statistics of mean, standard deviations, and percentage of values within 1, 2, and 3 standard deviations from the mean were calculated to evaluate normality. The goal of this normalization procedure was to establish a comparative scoring range for each of the models. The scores enable development of a combined overall model for filtering tweets relevant to the hurricane (as described in Section 2.4. Apart from the ratio, we also evaluated the F1-score (F1 = $2\frac{precision*recall}{precision+recall}$) for the model, where *precision* = $\frac{TP}{TP+FP}$, *recall* = $\frac{TP}{TP+FN}$, and *TP, FP, TN,* and *FN* represent the number of true positives, false positives, true negatives, and false negatives, respectively.

2.3.2. Image model

Supervised machine learning models were used to develop automated image classification of images in the Twitter dataset. The goal of this model is two fold: 1) to develop, a binary classifier capable of distinguishing hurricane-related images from the non-related ones, and 2) to then develop a multi-label annotator capable of classifying the hurricane-related images into one or more of three incident categories — 1) Flood, 2) Wind, and 3) Destruction.

A key hurdle in the approach was the lack of available labeled training data for supervised classification. We developed a web platform for image labeling for annotation by human coders. The platform took unlabeled images, and displayed them on a browser for human coders to annotate. Within the browser, the coder was asked the question of — *Does this image have any of the following — 1) Flooding, 2) Windy, and 3) Destruction?* An image is considered "Flooding" if there is water accumulation in an area of the image. An image is considered "Windy" if there are visual elements in the picture which show tree branches are moving in a direction or some objects that are flying or heavy rain visible in the image. An image is considered to have "Destruction" if there is damage to property, vehicles, roads, or permanent structures. An image can be in one or more of the previous classes. If an image has one of the codified classes, it is labeled as Irma related and if it does not have any of them the image is labeled as *not related*.

For the dataset, approximately 7000 images were labeled by 3 human coders/raters where the data was divided equally between them. Following codification resulting dataset had the following distribution — Related: 817/Not-Related: 6081 images, and Wind: 120 images; Flooding: 266 images; Destruction: 571 images. We also evaluated the inter-rater reliability using Light's Kappa [40] for 100 sampled images (with balanced distribution of related and not-related classes) that were labeled by all three coders. Agreement between all three coders for related versus not-related was at 0.77, and across tags Flooding - 0.88; Windy 0.27; and Destruction 0.78. This shows significant agreement among the coders on the labeling [41] other than the Windy tag (poor/chance agreement).

This annotated dataset was used to train deep learning models based on convolutional neural network (CNN) architectures. Convolutional networks have been widely used in large-scale image and video recognition [42]. CNN architecture consists of an input layer, an output layer, and several hidden layers in between. A hidden layer can be a convolutional layer, a pooling layer (e.g. max, min, or average), a fully connected layer, a normalization layer, or a concatenation layer. Within our approach, we evaluated three modern CNN architectures — 1) VGGNet, 2)ResNet, and 3)Inception-v3, and compared the performance of each model to its counterparts.

In VGGNet [42], the image is passed through a stack of CNN layers, where filters with a very small receptive field is used. Spatial pooling is done by five max-pooling layers, which is followed by convolution layers. The limitation of VGGNet is its large number of parameters, which makes it challenging to handle. Residual Neural Network (ResNet) [43] was developed with fewer filters and in turn has lower complexity than VGGNet. While the baseline architecture of ResNet was mainly inspired by VGGNet, a shortcut connection was added to each pair of filters in the model. In comparison, Inception-v3 [44] uses convolutional and pooling layers which are concatenated to create a time/memory efficient architecture. After the last concatenation, a dropout layer is used to drop some features to prevent overfitting before proceeding with final result. The architecture is quite versatile, where it can be used with both low-resolution images and high-resolution images, and can distinguish any size of a pattern, from a small detail to the whole image. This makes it useful in our application as the quality and type of image can vary widely due to disparate smart devices used by the Twitter population. The pre-trained Inception-V3 is trained on the Imagenet [45] dataset which consists of hundreds of thousands of images in over one thousand categories. The weights of this model are used as a starting point for training and fine tuned using our sample images. The approach takes advantage of transfer learning [46], where the classifier is able to initially learn features of physical objects in a wide variety of scenarios and then trained on specific observations within our data. This enables a more accurate and generalizable model.

Data augmentation methods were used to expand the number of training samples and therefore improve model accuracy. For example, additional training images are generated by rotating and scaling of the original images. This was done to balance the number of images of Irma related to the un-related ones. The resulting dataset consisted of approximately 6000 images in each class for the binary classifier, and approximately 2000 images in each class for the annotator model. The models were trained and testing using a 70-30 split on the dataset. For each model, performance scores (precision, recall, and F1) was recorded. Probability scores for each tweet image were then recorded for every class, which was further normalized using log-transform and re- scaled using min-max scaling to be used in the overall model.

2.3.3. User model

It is essential to quantify the authenticity of user accounts which have posted messages and images during a disaster event. For the purpose, our goal was to develop a scoring model which can provide continuous probabilistic measures of account authenticity.

Manually annotating user reliability in a large dataset such as Twitter is not practical. As we did not have a labeled dataset, our starting point was to consider the user "Verified" attribute within the tweets. The "Verified" attribute is annotated to user accounts which Twitter defines to be of public interest [47]. Within our dataset we had 94,445 non-verified users and 1692 verified users. Since Twitter's methodology for finding verified accounts is not public, we aim to develop a proxy automated model. The aim here is to create a model which can help identify users who are also likely to be accounts of public interest and authentic, but remain unverified. This can be used in conjunction with the Twitter "Verified" accounts to provide a comprehensive source of authentic accounts during a disaster event. Specifically, our approach provides the adjust the authenticity thresholds based on the continuous probabilistic scores of the model, which enables collection information from accounts which have not yet been verified by Twitter but have similar properties to that of a "Verified" account.

The automated model was developed based on supervised machine learning. Specifically, machine learning models [48] were developed for binary classification machine to predict the label user "Verified" (true/false) based on the features of tweets content (weblinks, hashtags) and its creator (account age, friends count, followers count, statuses count). Random Forest (RF) [49], Gradient Boosted (GB) [50], and Logistic Regression (LR) [51] classifiers were used to train and test the model.

RF is an ensemble model which consists of multiple decision trees trained on the data and their voting to determine the label class of an observation based on the features. A decision tree has a set of rules, when evaluated on an input, it returns a prediction of a class or a value. RF also returns the ratios of votes for each class it is trained on. A Gradient Boosted (GB) is also an ensemble model which builds decision trees leveraging gradient descent to minimize information loss. Similar to RF, GB also uses weighted majority vote of all of the decision trees for classification. In comparison, Logistic Regression (LR) is a non- parametric model which tries to find the best linear model to describe the relationship between independent variables and a binary outcome for classification.

The output of each of the trained binary models is a classifier capable of predicting if a user can be verified or not. The performance of the resulting model was evaluated using a 10-fold cross validation [52], with a 70-30 train test split used in each fold. Furthermore, grid search [53] was used on the best performing model for hyper-parameter optimization. Grid search takes in a set of values for each hyperparameter (e. g. number of trees in a forest, max depth of a tree, sample splits, max number of leaf nodes, etc.), folds number, and conducts a search using each possible combination of hyperparameters by evaluating them on a scoring metric such as F1-score. The final output of this model is a min-maxed log-transformed value of the probability scores. This was done to reduce the skewness in score distribution needed for the overall model (described later).

2.3.4. Text model

The goal of the text model was to delineate tweets with Irma related text from those addressing other topics. While generic search term such as the "Hurricane Irma" can provide a starting point, prior research [54–56] in the domain has shown that content organically develops to other words. An automated system trained on a large corpus to recognize context may improve the results, but this suffers from two significant pitfalls. First, training a learner on large bodies of text is costly from the perspective of computational overhead [57]. Second, the dynamic nature of discussions during a disaster, especially in a format as compact as Twitter, can alter the most likely interpretation of a word's meaning, resulting in false positives in the captured tweets [58].

In order to address the issues we developed a dynamic word embedding model which utilizes online learning to update its learned context. Specifically, we use a neural network based word embedding architecture - Word2Vec [59,60], which captures the semantic and syntactic relationships between the words present in tweets corpora. In the Word2Vec module, each word is evaluated based upon its placement among other words within a tweet. This target word, combined with its neighboring words before and after its occurrence in a given tweet, is then given to a neural network whose hidden layer weights correspond with the vector representing the target word. Once the vectors for each word are generated, the vectors can be compared based upon their cosine similarity. As two words get closer in similarity, the vectors representing those words will become closer within vector space; the angle internal to the vector will get smaller; and the cosine of this angle will get closer to, but not exceed, 1. As a result, the similarity in context between a word and its neighbors in vector space can be compared numerically by looking at the cosine of the internal angle formed by two word vectors [61].

Within our approach, tweets were parsed and grouped into 24-h segments, with primary testing done on the time period immediately before and after the initial landfall. Prior to training the model, tweets were first cleaned to eliminate punctuation, numbers, and extraneous/ stop words. Each tweet temporally isolated and parsed into token words, to create input vectors for training and testing of Word2Vec module. Four different formulas - 1) Cosine Similarity of Tweet Vector Sum (CSTVS) $1 - \frac{\alpha \cdot \sum_{i=1}^{k} \tau_i}{\|\alpha\| \left\| \sum_{i=1}^{k} \tau_i \right\|}$, 2) Dot Product of Search Term Vector and Tweet Vector Sum (DP) $\|\alpha\| \times \left\| \sum_{i=1}^{n} \tau_i \right\| \times \cos\theta$, 3) Mean Cosine Similarity (MCS) $\frac{1}{n} \sum_{i=1}^{n} \cos(\theta_{\alpha}^{\tau_i})$, 4) Sum of Cosine Similarity over Square Root of Token Count (SCSSC) $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \cos(\theta_{\alpha}^{\tau_i})$, were employed to score a tweet based upon its component word vectors. CSTVS is a programmatic implementation of the cosine distance formula [62] allows an efficient calculation of cosine distance. Cosine Similarity can be calculated by subtracting this value from 1. DP treats the sum of the vectors in a tweet as a vector itself $(\sum_{i=1}^{k} \tau_i)$, and calculating the dot product of this interpreted vector and the vector for the search term (α) returns a value that is proportional to the cosine similarity. MCS is the mean cosine similarity of the search term to all terms in a tweet, where n is the number of terms in the tweet. SCSSC is similar in function to the MCS, where it reduces the impact of a shorter tweet by dividing by the square root of the count of tokens in a tweet (n). All formulas return a scalar score for a tweet - search term similarity match.

In order to evaluate the model, the codified data set of 19,000 tweets were used. The codification was done by a single human coder and a sampled set of tweets (100 with balanced distribution) was verified by two additional coders to access the inter-rater reliability. Tweets were labeled to be Irma related if matched the following criterion — 1) *Explicitly contains references to "Irma" or "Hurricane"*; 2) *Contains current meteorological data, such as wind speed, rainfall levels, etc.;* 3) *Refers to weather events such as storm, flood(ing) and rising water, rainfall, tornado, etc.;* 4) *Describes the aftermath of extreme weather: trees down, power out, damage to buildings or construction, etc.;* 5) *contains references to emotional states exacerbated by the weather: worrying about shelter, concerns for safety, pleas for help, etc.;* 6) *Lists availability or absence of necessities: shelter, water, food, power, etc.* A message was labeled not related if it met following criterion — 1) *mentioning a location absent any of the above content;* 2) *Containing an attached picture that may be Irma related, but no additional text;* 3) *Expressing emotions about the state of an event, but its connection to weather is ambiguous, i.e. a sporting event canceled, but no explanation as to why;* 4) *Expressing emotions about a person's condition, but its connection to weather is ambiguous: for ex:* "I hope @abc123 gets better soon!". The resulting dataset had 8296 tweets related to the Irma and 10,792 tweets not related. The inter-rater reliability of the codified messages using Light's Kappa metric was at .69, suggesting significant agreement between coders [40,41].

This dataset was then used to evaluate the aforementioned formulas for different thresholds of the scores by analyzing the ratio of correctly classified tweets by the model. Hyper-parameters of the Word2Vec model were also tuned using the labeled tweets. The parameters selected for testing were context word window sizes from 1 to 10 words on either side of the target word; hidden layer dimensionality in 50D increments from 50D to 500D; minimum word occurrence from 0 to 9; negative sampling from 0 to 9 words. The cross product of the values contained in these ranges were used as the testing set of tuples for the training operations. For each set of parameters, the NN was trained through varying epochs, and the resultant word embeddings used in conjunction with the four scalar formulas to calculate scores for each tweet. The scores for each iteration were min-max scaled for the time delta, and the AU-ROC calculated based upon the thresholds of the scores in relation to the human-coded tweets.

2.4. Overall model

Following the creation of individual models, we combined the results of each into a single overall model (Fig. 2) which consists of three distinct stages - 1) Metadata extraction, 2) Filtering, and 3) Visualization of filtered tweets. For the first stage, the input is a tweet as a data-point. The metadata extraction stage mines the relevant attributes (image, geolocation, user, text) needed for the individual models of 1) Geospatial, 2) User, 3) Image, and 4) Text analysis.

The results of the individual models are then combined in the second stage of filtering, where the normalized scores (decision score ranging from 0 to 100) for each models are combined at different thresholds to filter the relevant Twitter messages for Hurricane Irma. Any tweets without images are assigned an imgScore=0, this allows users to view messages which contain images by setting the threshold to be imgScore> 0. The flexibility of the approach is in its ability to select different thresholds for respective models. This allows for a more generalizable model where a user can choose different set of thresholds for disparate disaster events. A logical AND operation is used to obtain messages which pass all of the thresholds for each of the individual models. Specifically, a datapoint can only pass the filtering stage if all of its individual model scores are greater than or equal to the thresholds set.

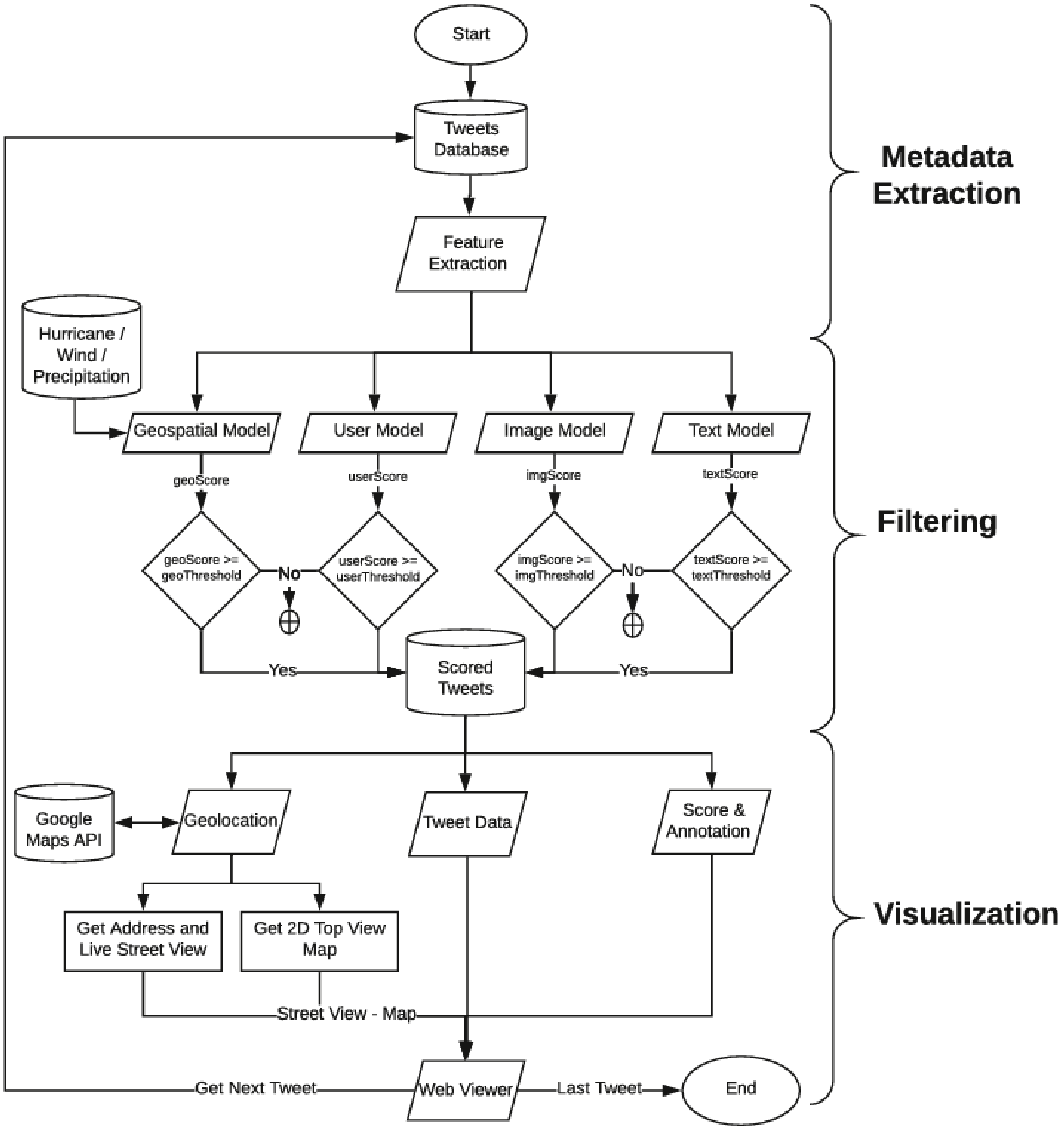


Fig. 2. Overall information flow model. The Metadata Extraction stage develops variables from the raw Twitter data, Filtering stage utilizes the developed 1) Geospatial, 2) User, 3) Image, and 4) Text analysis modules to score tweets, and the Visualization stage is used to observe the at location posted image along with Google Street View.

The filtered data are then stored in a database (Scored Tweets), where each datapoint can then be viewed on a visualization platform. The visualization platform extracts the location information from each datapoint (Geolocation), which is then cross-referenced with Google Maps API to provide three attributes — 1) Google Street View [63,64], 2) Physical address, and

3) A 2D top down view of the map at the location. These attributes (Street View - Map) along with the Tweet Data (text of the tweet, date-time, user, image, etc) and Score & Annotation information (P(*Related/NotRelated*)and P(*Tag*), where P is the probability and *Tag*∈{*Flooding,Windy,Destruction*}) is then displayed on a web viewer. This presents an easy to use interface to view and visualize the messages for situational awareness.

3. Results

3.1. Geospatial

Preliminary exploration of the sensor readings for precipitation and wind speed with relative distance from the eye of the hurricane are shown in Fig. 4. Precipitation decreases exponentially farther away from the eye of the hurricane, measuring 5–20 inches. Median wind speeds have their peak around 300 miles from the eye of the hurricane.

Nine different geospatial models were developed and compared for their performance to filter Irma related tweets. Specifically, for each model the results calculated ratio of Irma related tweets, i.e. number of Irma related tweets/total number of tweets, at different thresholds between 0 and 1 (all values were min-maxed for normalization). Irma related tweets were identified by codification of 19,000 messages by human coder (annotation criteria described in Section 2.3.4). Fig. 3, compares the cumulative distribution function (CDF) plots between the each of the functions within a subplot. The plots (a, b, c) further compare the results between - a. Min-Max Normalization, b. Log ($\log_{10}()$), and c. Box-Cox ($\gamma()$)) transformation scores (see Table 1).

As observed, the results of the Log and Box-Cox transformations show a wider distribution of the ratio in comparison to the Min-Max normalized values, across the different thresholds. The results are also confirmed by the Shapiro-Wilks test (Table 2) where the Log and Box- Cox transformed models have higher scores, suggesting a more normal distribution of the results than the non-transformed ones. Based on the test the top five functions identified were – ($\gamma\left(\frac{wind}{\sqrt[3]{distance}}\right), \gamma\left(\frac{wind*rain}{distance}\right), \log_{10}\left(\frac{wind*rain}{distance}\right), \gamma\,\frac{wind*rain}{\sqrt{distance})}$, and $\log_{10}(\frac{wind*rain}{\sqrt{distance}})$. Each of the models were in very close proximity to the scores observed in the test.

Additional analysis was conducted to observe the statistical properties of the top five models. Fig. 5 shows the CDF and F1-Scores for each of these functions. Table 2, show the general statistical properties. Out of the five, $\log_{10}(\frac{wind*rain}{\sqrt{distance}})$ was chosen as a final model function, based on its mean being the closest to 0.5.

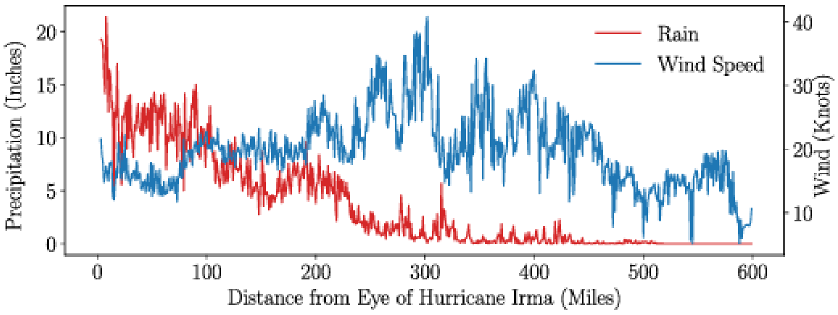
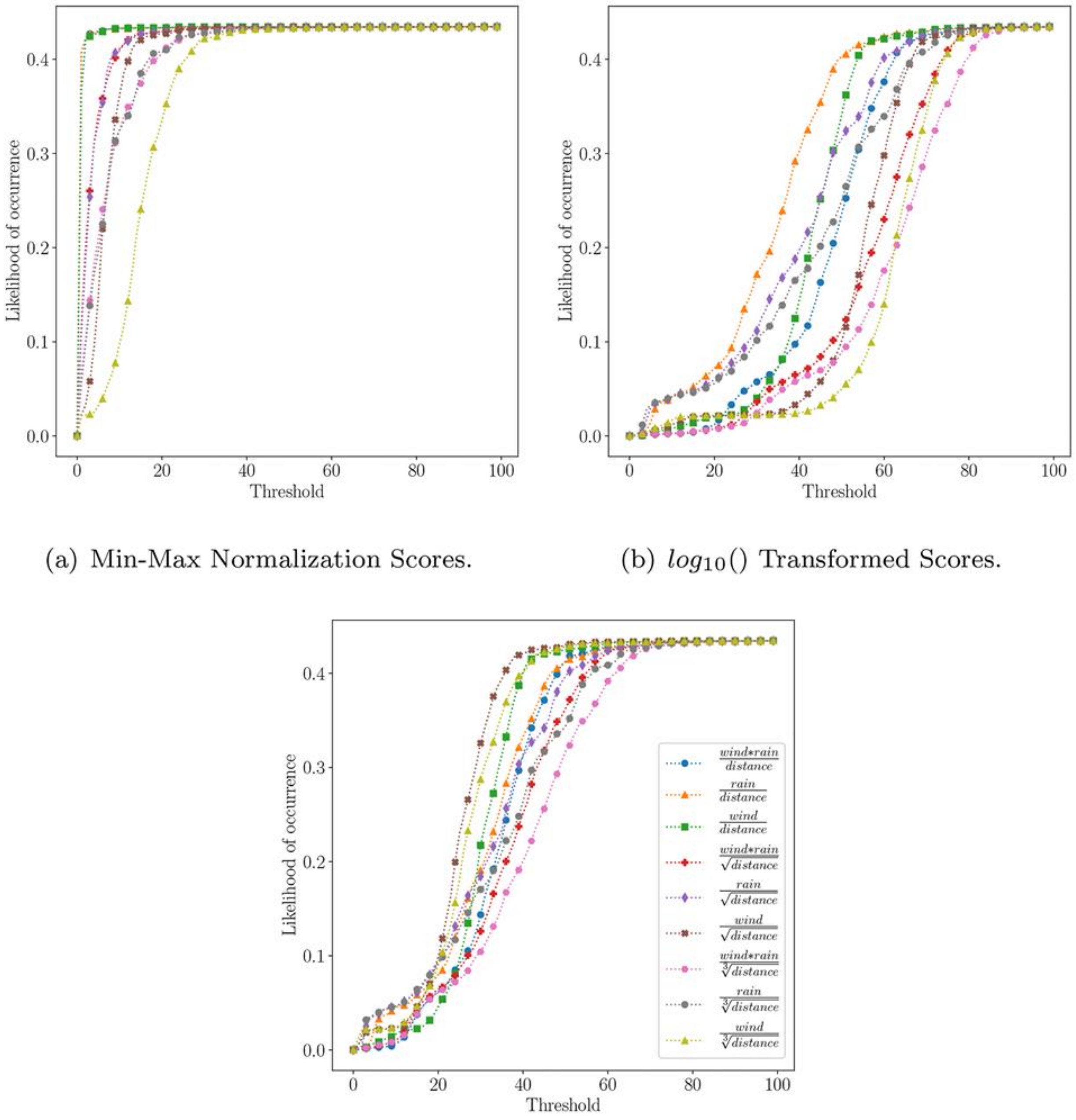


Fig. 3. Precipitation and wind speed in relation to the distance from Hurricane Irma's eye.

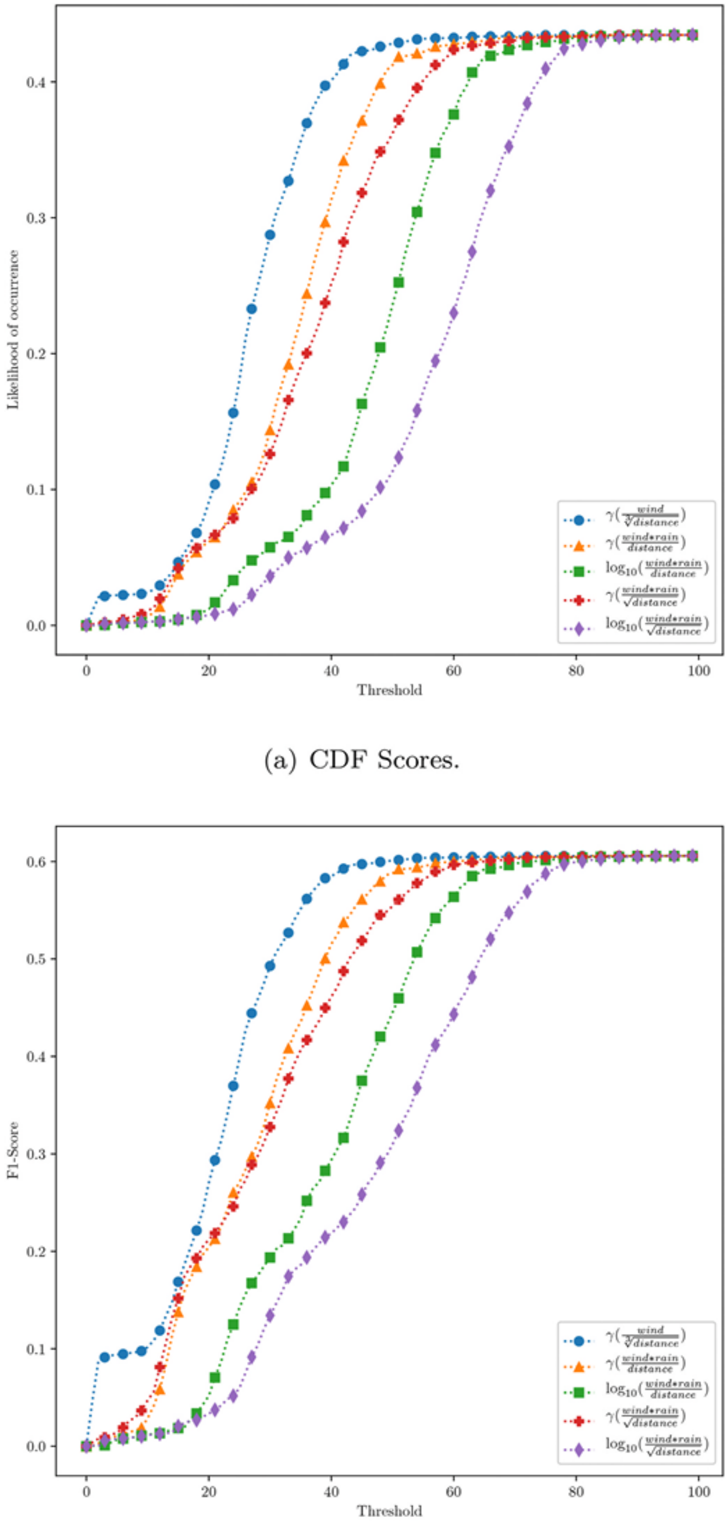


(a) Min-Max Normalization Scores.

(b) $log_{10}()$ Transformed Scores.

(c) Box-Cox Transformed Scores.

Fig. 4. Cumulative Distribution Function (CDF) for Min-Max Normalization, Log, and Box-Cox transformed geospatial scores for the nine models. The common legend of all three figures is shown in Figure c.



(a) CDF Scores.

(b) F1 Scores.

Fig. 5. Cumulative Distribution Function (CDF) and F1-Scores for the top five geospatial models.

Table 1. Shapiro-Wilk statistics value for all the models. Top 5 values highlighted.

Model	Normalization method		
	$\frac{X - \min(X)}{\max(X) - \min(X)}$	$log_{10}()$	$\gamma()$
$\frac{wind * rain}{distance}$	0.07	**0.99**	0.99
$\frac{rain}{distance}$	0.06	0.92	0.93
$\frac{wind}{distance}$	0.13	0.95	0.97
$\frac{wind * rain}{\sqrt{distance}}$	0.53	**0.98**	**0.98**
$\frac{rain}{\sqrt{distance}}$	0.51	0.88	0.89
$\frac{wind}{\sqrt{distance}}$	0.88	0.88	**0.98**
$\frac{wind * rain}{\sqrt[3]{distance}}$	0.65	**0.98**	**0.98**
$\frac{rain}{\sqrt[3]{distance}}$	0.65	0.85	0.86
$\frac{wind}{\sqrt[3]{distance}}$	0.96	0.85	**0.99**

3.2. Image classification

The performance of various image classifiers are shown in Table 3. In the first stage of classification, which uses a binary classifier distinguish hurricane and non-hurricane related images, the Tuned Inception V3 architecture performed the best with an overall F1-score of 0.962. Fig. 6, shows the comparative AU-ROC curves for the different models. Between the classes, the Tuned Inception V3 model also performed well with an F1-score of 0.959 for class 1 (hurricane related) and 0.965 for class 0 (non-hurricane related) images.

The performance of various image classifiers are shown in Table 3. In the first stage of classification, which uses a binary classifier distinguish hurricane and non-hurricane related images, the Tuned Inception V3 architecture performed the best with an overall F1-score of 0.962. Fig. 6, shows the comparative AU-ROC curves for the different models. Between the classes, the Tuned Inception V3 model also performed well with an F1-score of 0.959 for class 1 (hurricane related) and 0.965 for class 0 (non-hurricane related) images.

The performance of various image classifiers are shown in Table 3. In the first stage of classification, which uses a binary classifier distinguish hurricane and non-hurricane related images, the Tuned Inception V3 architecture performed the best with an overall F1-score of

0.962. Fig. 6, shows the comparative AU-ROC curves for the different models. Between the classes, the Tuned Inception V3 model also performed well with an F1-score of 0.959 for class 1 (hurricane related) and 0.965 for class 0 (non-hurricane related) images.

Table 2. General data statistics for top 5 models. $\log_{10}(\frac{wind*rain}{\sqrt{distance}})$ General data statistics for top 5 models.

Model	Data Statistics			
	Shapiro-Wilks	Standard Deviation (σ)	Mean μ	% of Data within -1 ≤ σ ≤ 1
$\gamma\left(\frac{wind}{\sqrt[3]{distance}}\right)$	0.99	0.12	0.28	0.66
$\gamma\left(\frac{wind * rain}{distance}\right)$	0.99	0.14	0.38	0.65
$\log_{10}(\frac{wind *\ rain}{distance})$	0.99	0.14	0.39	0.65
$\gamma\left(\frac{wind * rain}{\sqrt{distance}}\right)$	0.98	0.16	0.43	0.64
$\log_{10}\left(\frac{wind *\ rain}{\sqrt{distance}}\right)$	**0.98**	**0.16**	**0.46**	**0.64**

3.3. User

The F1-score of the Random Forest (RF), Gradient Boosted (GB), and Logistic Regression (LR) models of the models trained on predicting user verification were recorded at 0.97, 0.92, and 0.88 respectively. Fig. 8 shows the comparative AU-ROC scores of the different models, where the RF classifier is able to outperform the rest of the models. The best performing RF model was developed by using a grid search approach, where multiple model parameters (number of estimators, depth, leaf splits, etc.) were evaluated. The resulting model had a precision, recall, and AU-ROC were observed to be 0.96, 0.98, and 0.99 respectively.

The classifier was balanced in its prediction accuracy in both verified (class 1) versus non-verified (class 0) users (Fig. 10). The output probability values of the binary model were further min-maxed to a threshold score between 0 and 100. The resulting normal distribution had a mean of 50.56, a median of 66.26, and a standard deviation of 39.69.

3.4. Text

The results of the text analysis module were based on the binary categorization of the tweets codified as ‘irma related’ (class 0) or ‘non irma related’ (class 1). Evaluation of the four different resulted in the F1- scores of .6553 - MCS, .7824 - DP, .7049 - CSTVS, and .7347 - SCSSC. We observe the dot product between search term vector and tweet vector sum (DP) gives us the best result. Fig. 9 shows the AU-ROC curves comparing the different formula performance in the analysis.

Each model was further evaluated to identify the best set of parameters. Within the analysis we found the DP formula was still the best performing with a word window size of 1, hidden layer dimensionality of 150, a minimum word count of 5, a negative sampling value of 1, and training the Word2Vec model through 25 epochs. The resulting normal distribution had a mean of 24.73, a median of 21.64, and a standard deviation of 14.05.

Table 3. Performance comparison of deep-learning models (Inception-V3, VGGNet, ResNet, and Tuned Inception-V3) for binary classification and multi-label annotation.

Model	Performance Measures					
	Binary Classifier			Multi-Label Annotator		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
VGGNet	0.88	0.87	0.88	0.70	0.60	0.64
ResNet	0.88	0.89	0.89	0.68	0.61	0.64
Inception-V3	0.89	0.88	0.88	0.75	0.72	0.73
Tuned Inception-V3	**0.96**	**0.95**	**0.95**	**0.90**	**0.92**	**0.91**

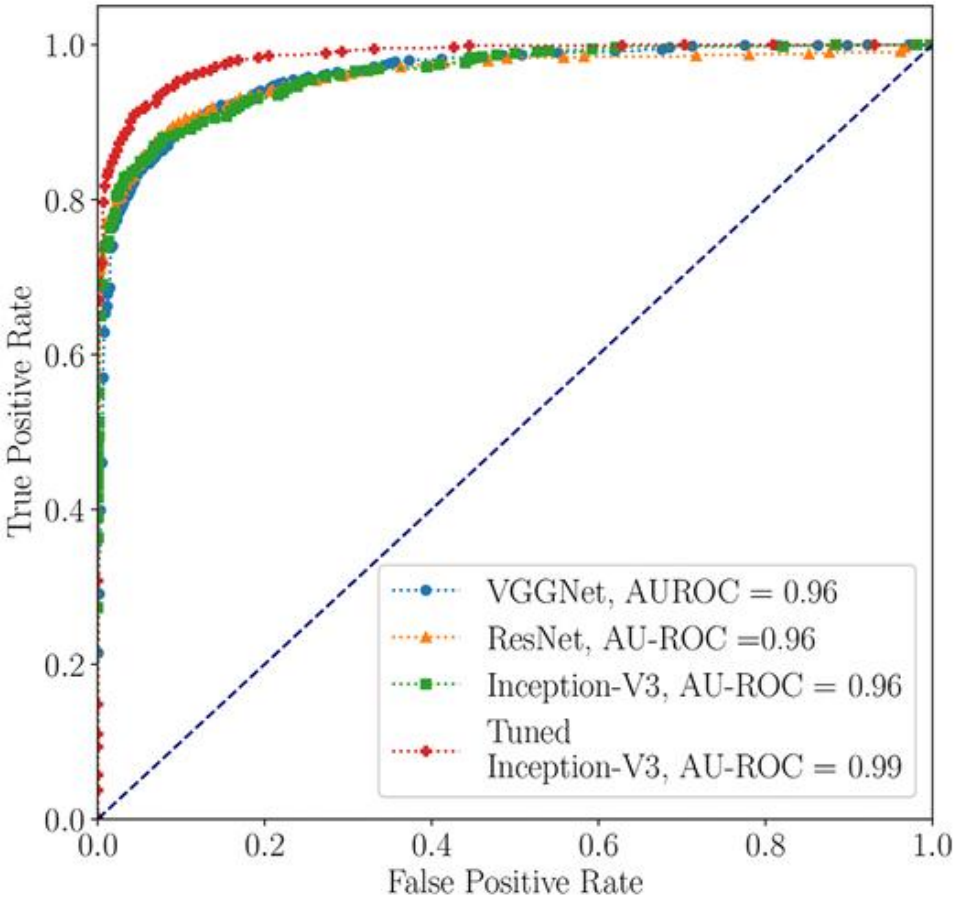


Fig. 6. Area Under - Receiver Operating Characteristics (AU-ROC) Curves for V3, VGG net, ResNet architecture and Tuned Inception V3 models for binary classification of images (hurricane related versus non-hurricane related).

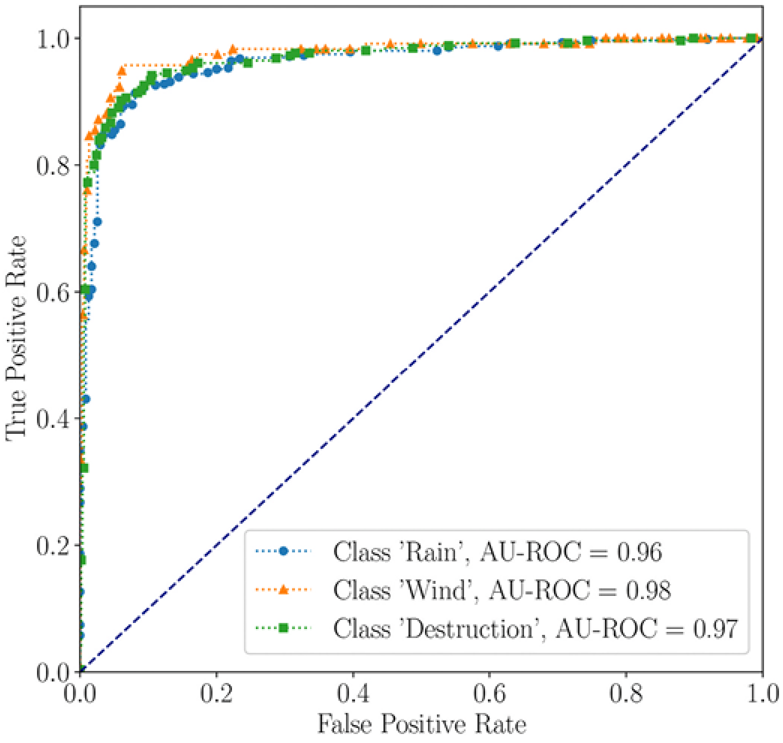


Fig. 7. Area Under - Receiver Operating Characteristics (AU-ROC) Curves for Tuned Inception V3 model for multi-label annotation for images - 1) 'Flood', 2) 'Wind', and 3) 'Destruction'.

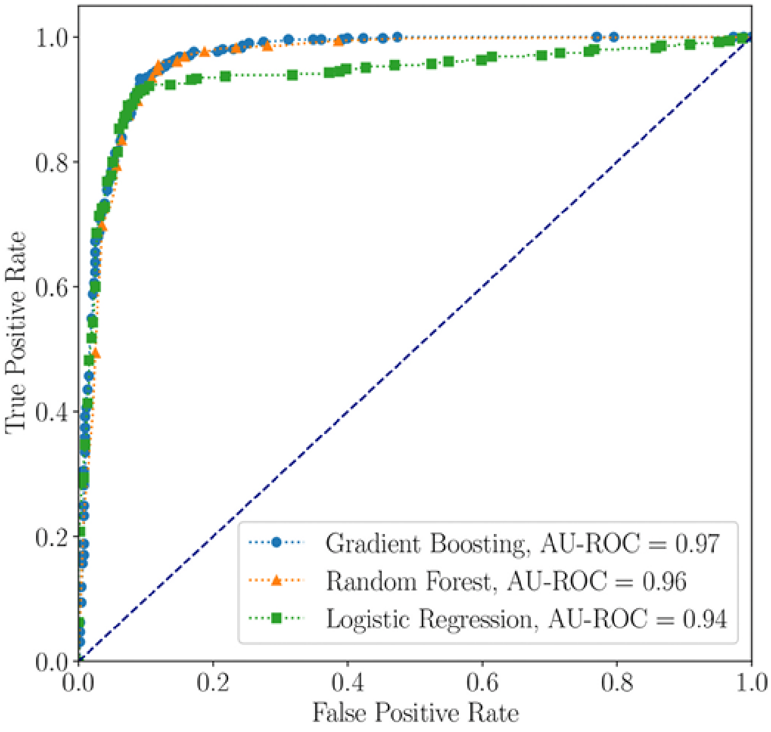


Fig. 8. AU-ROC Curves for Random Forest, Gradient Boosted, and Logistic Regression Classifiers in predicting Verified users.

4. Discussion

We address each individual model separately before discussing the final combined model and providing limitations/future directions for this work.

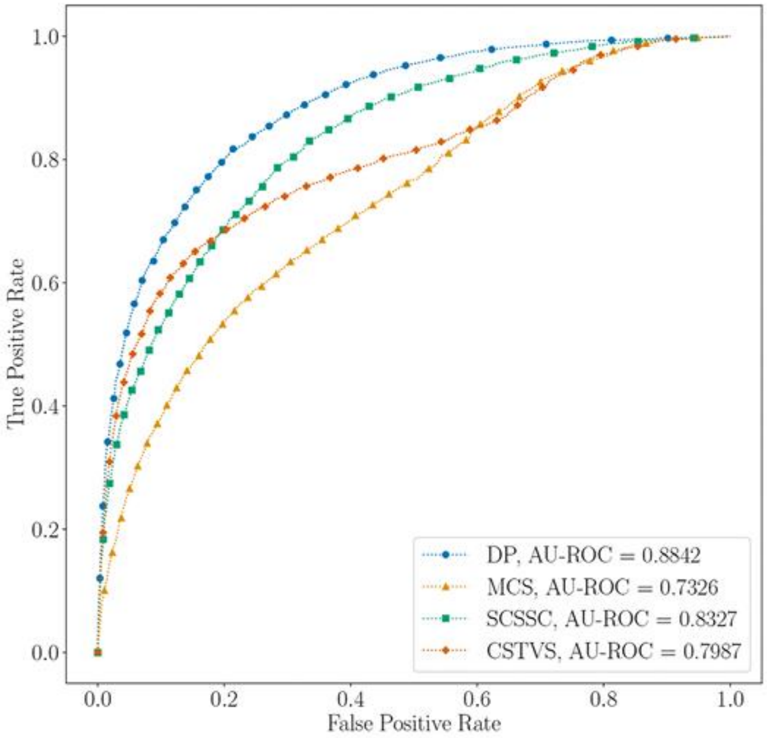


Fig. 9. AU-ROC Curves for text — 1) Cosine Similarity of Tweet Vector Sum (CSTVS), 2) Dot Product of Search Term Vector and Tweet Vector Sum (DP), 3) Mean Cosine Similarity (MCS), 4) Sum of Cosine Similarity over Square Root of Token Count (SCSSC).

4.1. Individual models

4.1.1. Geospatial

The geospatial models developed in the study provide a measure of relevance to a tweet by including the forcing sensor data (wind speed, precipitation, and distance from the eye of the hurricane). The best performing function of $log_{10}\left(\frac{wind * rain}{\sqrt{distance}}\right)$ combines the values into a single normalized score which can be used to weight a geographic/ sensor relevancy factor for any tweet. More specifically, the function helps us identify Twitter messages at locations which are in close proximity to the hurricane forcing and have observed increased amount of precipitation and wind speed. As seen in the results, the chosen $log_{10}\left(\frac{wind * rain}{\sqrt{distance}}\right)$ was the closest to a normally distributed function. This allows for a greater granularity on threshold cutoff points in comparison to other functions, leading to a fine-grained control over filtering based on the geographic relevance of the tweets. The statistical properties of the function also enables analysis of confidence intervals which can be used to ascertain the reliability of a message within the context of sensor data. In other words, tweets with anomalous sensor readings can be easily identified, leading to more reliable mining of messages related to the disaster event.

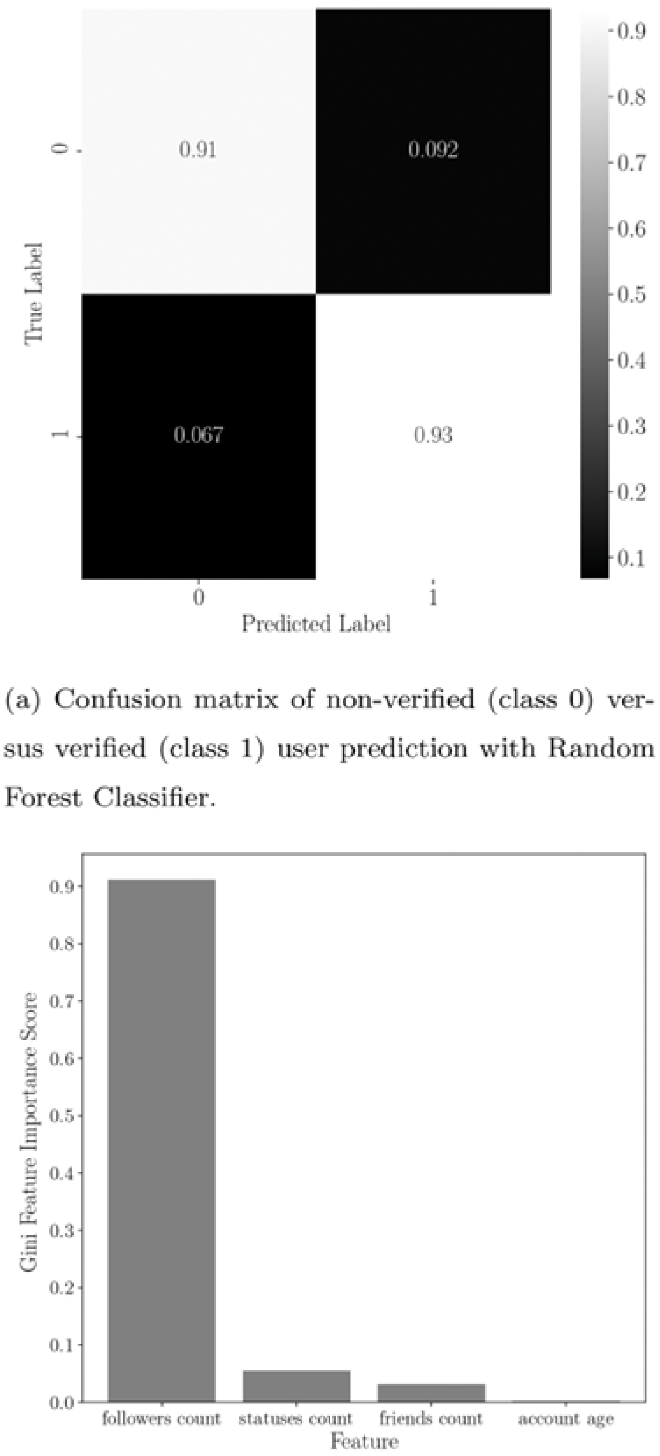
We envision that filtering tweets using their geospatial information relative to storm position and also environmental factors can help isolate tweets from heavily impacted locations. By examining locations close to the storm, with high wind gusts, or heavy precipitation allows users to quickly examine locations that might be expected to show the most severe impacts from storm events.

4.1.2. Image

Comparing the performance of the CNN architectures (Inception V3, VGG, ResNet, and Tuned Inception V3) for binary classification (hurricane and non-hurricane related), we observe that the Tuned Inception V3 model (F1-score 0.95) has almost a 6–7% accuracy gain over others. In comparison to the VGG and ResNet architectures, the Tuned Inception V3 larger number of parameters which can be trained to observe the nuances between the images. While the base Inception V3 classifier contains the same number of parameters, re-tuning the weights to our training sample of images improved its accuracy considerably for the binary classification. This can be attributed to the pre-training and transfer learning of the model, where it already had prior weights based on classification of physical objects, and our image data tuned it further for disambiguating physical and non-physical scenes.

We do observe a slight performance decrease (F1-score 0.91) of the architecture trained on the multi-label annotation of the images. This can be attributed to the limited number of training samples that were available to the classifier. The complexity of the images in the samples further degrades the performance, for example, images of lakes and sea water are not much different from images of flooding.

Prior research in the area of automating image analysis (using machine learning) from social media has primarily focused on quantifying the level of damage in disaster situations [65–67]. Our approach uses a dual stage model, where the first stage is responsible for increasing the quality of images by filtering out the non-relevant/non-physical images. The output is then fed into the second stage for categorization into different groups based on situational conditions (flooding, wind, and destruction). While prior studies have looked at disambiguating "fake/altered" images [68,69], they are based on analyzing the content of the tweet along with user reliability measures for training machine learning models. Within our approach we only utilize the image features for the training our models. The output is image scores are based on normalized probability values, which can be used for threshold cutoffs, where setting a high threshold will only mine the most hurricane related images. The second stage then annotates the images for further filtering of images based on the needs of the domain. Filtering images permits users to quickly focus on a small subset of visual information that is presumed to be most valuable for storm impact assessment, compared to needing to scroll through many images to find useful information.



(a) Confusion matrix of non-verified (class 0) versus verified (class 1) user prediction with Random Forest Classifier.

(b) Relative importance between features (using information gain) for prediction verified users using Random Forest Classifier.

Fig. 10. Performance of the user Random Forest Classifier in the binary classes and feature importance metrics.

4.1.3. User

Prior studies [70–73] focused on identifying incorrect/fake/altered information in social media have established the source of information (social media user) as a key component. A large proportion of the studies [74–77] have been based on developing machine learning approaches towards detection of “bots” or fake user accounts [78] on social media. For example [79], identify the credibility of the user as an important element in mining good quality situational awareness information from social media. Within our approach, we leverage prior work done in the field by identifying the user features of account age, status count, number of followers, number of friends, existence of url links, number of hashtags, existence of images, retweets, geolocation, and message frequency in training our machine learning models.

Comparing the results between the parametric (Logistic Regression) and the non-parametric ensemble models (Random Forest and Gradient Boosting), we observe the ensemble models are able to outperform by a margin of 4–7%. The developed Random Forest model has a very high accuracy (F1-score 0.97, AU-ROC 0.99) in disambiguating between verified and non-verified users. While the ratio of the number of verified versus non-verified users was imbalanced (approximately 1:100) in our data, the developed RF model is able to accurately distinguish between the classes as shown by the confusion matrix (Fig. 8).

Further analyzing the RF model, we calculated the average decrease in Gini impurity/information gain (entropy) among all estimators to observe the importance of features. Specifically, as estimators are developed on a subset of features, the decrease in information gain across a subset of features can be used to infer the relative importance of features. Fig. 9, shows the relative importance of four features (rest where too low to observe), where the number of followers, status, and friends, along with the account age are the top features which affect the decision of the model towards the credibility of a Twitter user in our data.

The analysis of features within our model shows similar feature importance measures to that used in prior research to identify reliable information sources [79]. However, our approach provides a more generalized model where a thresholding on probability scores can be used to select user sources based on needs of a specific event. The approach is also dynamic where a model can be quickly retrained using the available “Verified” tags instead of manually re-annotating accounts. This prevents temporal dilation of features where a model trained on an older labeled dataset cannot perform as well due to the changes in account statistics over time.

4.1.4. Text

With the observed dot-product based model performing the best with the F1-score analysis, we applied the model towards an hourly aggregated corpus within our data. Specifically, when the corpus was confined to the tweets from a single hour, the vector representations of word embeddings were only influenced by the contexts derived from that hour. Words would have a unique vectorization specific to that hour, and relationships between words were dependent on the context interpretations within that time. The cosine similarity of two terms could be calculated for this duration, and words with the highest scoring cosine similarity to a term would indicate an observed relationship that was finite within the timeframe. In short, two words could be similar in 1 h, and completely different the next, depending on the content of the tweets at the time.

Table 4 shows the output of the DP model for the hourly aggregated tweets. Prior to landfall (time 13:00), we observe mentions of the “storm”, “wind”, “eye”, “ese” (East-South-East), “e” (East), etc, having prominence in the top 20 words as identified by the DP model to be semantically similar to search term “irma”. There is consistency in the thematic representation where these words did occur across the 6 h prior to the hurricane. During the window of the hurricane Irma’s landfall we observe “shelter”, “#hurricaneirma”, “eye”, “landfall”, “help”, “plea”, etc., as the most related terms to “irma”. After the hurricane the context of the “irma” changes to reflect more help/rescue/concern words where “shelter”, “safe”, “check”, “food”, “power”, etc. become the most prominent words.

The results show that the word embedding based dot product model is capable of identifying tweets which are most relevant to the search/ seed term. This is highlighted by the example of the term “ese”, which when taken by itself, might reference an informal Spanish colloquialism for “man”. When interpreted within the hourly-divided corpora within this dataset, it takes on a different semantic interpretation. For the tweets occurring within each of the 4 h immediately preceding landfall, “ese” is in the top twenty most related terms to “irma”, and does not appear in the hourly lists following. Looking at the terms related to “ese” it can be determined that this refers to the abbreviation for East-South- East, likely referencing the direction from which the hurricane approached. After landfall, this term was no longer as relevant, and therefore less likely to appear as a related term.

4.2. Overall model

In the overall model, the number of possible combinations for the thresholds is large (at 1004), where each of the four models can have a value ranging from 0 to 100. A cumulative distribution plot (CDF) was used to analyze the percentage of data-points passing the thresholds set for each of the models. Fig. 11 shows the comparative analysis of each of the model, where the curves are inversely proportional to the thresholds indicating a decrease in the percentage of tweets passing higher thresholds. Within the analysis, low thresholds are representative of more reliable sources and related contents, resulting in a low percentage of overall tweets passing through the filter. Similarly, at the threshold of 100, all tweets pass the filter providing complete access to all data.

The CDF plot also highlights the comparative performance of various models, where the text based filter includes a higher percentage of tweets at lower thresholds while the user verification filter includes most users at higher thresholds. Image classification also results in a similar performance to user verification (including most users at higher thresholds), whereas filtering based on geospatial scores filters more linearly. We observe high quality results (low false positives) at a likelihood occurrence of 0.6. by setting initial thresholds to 30 for text, 50 for geospatial, and 85 for both image and user scores. These recommended thresholds for Hurricane Irma provide a baseline for comparison with different events, and for hurricanes in different locations.

4.3. Limitations and future work

Our current work explores the utilization of multiple modalities present in social media data to filter hazard event related information. We acknowledge certain limitations of this approach. Our approach is to cumulatively evaluate the operation of all sub-models in capturing the messages. As a result, we focused in this work on tweets that have all attributes: geolocation, text, and image (note all tweets have user attributes). However a smaller overall model with specific combinations of the sub-models can be used in certain conditions. For example, researchers who are interested in just messages with text can use an overall model that excludes the image sub-model and subsequently not filter based on a threshold for images.

Furthermore, our models are evaluated using the data from a single event — (Hurricane Irma) and a single location (Florida, USA). As a part of our future effort we plan to extend this framework to other hurricane events (and locations), such as, Maria, Harvey, and Florence, along with application of the approach to other disaster scenarios, such as fires, earthquakes, floods, etc. to aid in understanding the filtering step and thresholds in other contexts. Each event will likely have different specifications on the quality of data that needs to be extracted, for which we need to cross evaluate the approach against various events to provide recommendations for thresholds to be used for different disaster categories.

Our approach can operate as a primary filtering mechanism for additional analysis to extracting information during a disaster event. Additional models which help with categorization of messages, such as, disaster damage quantification, information, requests of help, resources offerings, organizing efforts, etc., can be implemented to extract higher level information from the data.

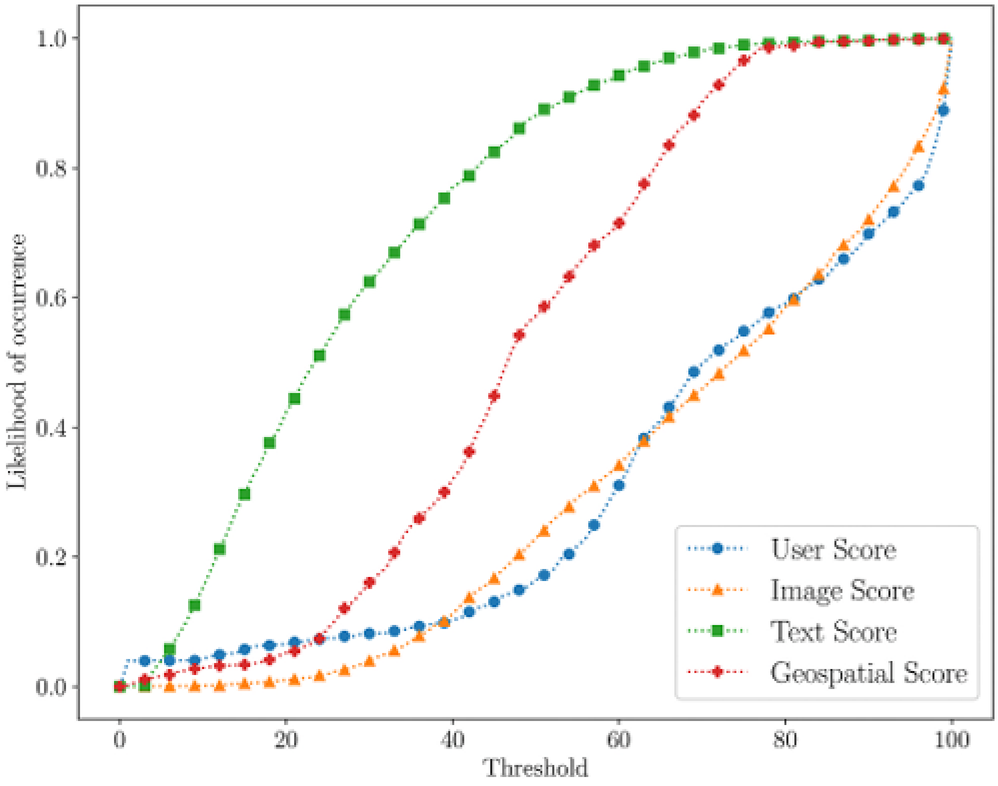


Fig. 11. CDF of Overall Model and percentage of tweets passing different model thresholds.

Table 4 Hourly aggregate of top 20 semantically related to terms to " Irma" , for 6 h prior and after landfall. The words have been stemmed to their root. #hirma denotes the hashtag #hurricaneirma used in the tweet. Colors indicate similar terms across the different time windows of the hurricane.

Word Rank	**Time**												
	7:00	**8:00**	**9:00**	**10:00**	**11:00**	**12:00**	**13:00** **Irma Landfall**	**14:00**	**15:00**	**16:00**	**17:00**	**18:00**	**19:00**
1.	sleep	last	ese	ese	tampa	tampa	**shelter**	shelter	shelter	shelter	#hirma	hit	tampa
2.	offici	outsid	outsid	tri	yet	time	**first**	whole	tampa	want	outsid	safe	eye
3.	need	sleep	moder	help	time	check	**wait**	tampa	beauti	time	food	outsid	bay
4.	ese	ese	beauti	yet	see	good	**get**	open	first				
5.	want	e	sleep	outsid	eye	tri	**tri**	check	#hirma	tampa	safe	updat	time
6.	hit	#key	nation	eye	friend	might	**make**	open	prep	get	watch	hurrican	wait
7.	#key	wind	e	moder	first	night	**could**	hit	food	guess	time	make	hit
8.	wind	tropic	need	sleep	night	first	**made**	get	come	hurrican	peopl	prayer	outsid
9.	tropic	good	wind	heavi	last	close	**see**	friend	watch	last	know	first	#hirma2017
10.	see	beach	fuck	wellington	close	coffe	**eye**	read	yet	check	see	get	make
11.	much	#irma	#sfltraffic	wind	#traffic	friend	**#hirma**	world	time	hit	love	wait	food
12.	#irma	florida	pleas	fuck	strong	help	**world**	safe	sleep	peopl	power	everyon	us
13.	beach	storm	storm	tropic	outsid	last	**night**	good	ride	come	gonna	check	shelter
14.	florida	#mfl	beach	good	make	follow	**peopl**	time	first	eye	#irma	see	get
15.	storm	aso	peopl	see	well	outsid	**close**	come	get	friend	still	home	last
16.	know	lauderdal	#irma	pleas	want	make	**outsid**	make	check	food	hurrican	power	point
17.	#mfl	power	florida	storm	phone	sleep	**help**	first	go	day	#nfl	#hirma	safe
18.	power	mesonet	f	flood	sleep	strong	**landfal**	beauti	know	see	make	okay	open
19.	call	rain	rain	beach	hit	#irmageddon	**come**	wait	open	make	home	watch	alway
20.	aso	safe	mesonet	rain	florida	open	**pleas**	home	tri	way	want	yet	video

5. Conclusion

In this study, a multimodal filtering approach was developed and evaluated to extract and subset geocoded images posted on Twitter within the context of Hurricane Irma. Our prototype model consisted of four sub-models: geospatial, image classification, user credibility, and text analysis. Each sub-model returned a score in the range of 0–100 and allowed for user-defined filtering based on bespoke thresholds. Each of the four models aim to filter information about reliability, information consistency, and overall usefulness of the message. This single combined model shows potential for application in disaster and emergency contexts, allowing users to quickly search and filter for relevant geolocated tweets.

Data and codes availability statement

The data and the codes used in the research are available on Figshare: https://figshare.com/s/235146fc2d6de33654f3 [80].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank UNCG undergraduate students Kaitlyn Jessee and Elaina Kauzlarich for classifying the images associated with the tweets. This study was supported in part by the National Science Foundation (Awards# CMMI-1541136 and # SES-1823633), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (Award #P2C HD041025), DoD/DARPA (R0011836623/ HR001118200064), seed grant initiatives from UNC Greensboro and Penn State's Social Science Research Institute, Population Research Institute, and Institute for CyberScience, and an Early-Career Research Fellowship from the Gulf Research Program of the National Academies of Sciences, Engineering, and Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Gulf Research Program of the National Academies of Sciences, Engineering, and Medicine.

References

1. T. Knutson, S. J. Camargo, J. C. L. Chan, K. Emanuel, C.-H. Ho, J. Kossin, M. Mohapatra, M. Satoh, M. Sugi, K. Walsh, L. Wu, Tropical Cyclones and Climate Change Assessment: Part II. Projected Response to Anthropogenic Warming, Bull. Am. Meteorol. Soc. doi:10.1175/BAMS-D-18-0194.1. URL https://journals.ametsoc.org/doi/10.1175/BAMS-D-18-0194.1.
2. H.R. Moftakhari, A. AghaKouchak, B.F. Sanders, D.L. Feldman, W. Sweet, R. A. Matthew, A. Luke, Increased nuisance flooding along the coasts of the United States due to sea level rise: past and future, Geophys. Res. Lett. 42 (22) (2015) 9846–9852,

doi:10.1002/2015GL066072, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL066072.

3. B. Neumann, A.T. Vafeidis, J. Zimmermann, R.J. Nicholls, Future coastal population growth and exposure to sea-level rise and coastal flooding - a global assessment, PloS One 10 (3) (2015), e0118571 doi:10.1371/journal. pone.0118571, https://journals.plos.org/plosone/article?id=10.1371/journal. pone.0118571.
4. E.D. Lazarus, P.W. Limber, E.B. Goldstein, R. Dodd, S.B. Armstrong, Building back bigger in hurricane strike zones, Nat. Sustain. 1 (12) (2018) 759–762, doi:10.1038/s41893-018-0185-y, https://www.nature.com/articles/s41893-018-0185-y.
5. B. De Longueville, R. Smith, G. Luraschi, ”OMG, from Here, I Can See the Flames!”: a Use Case of Mining Location Based Social Networks to Acquire Spatio-Temporal Data on Forest Fires, 2009, pp. 73–80, https://doi.org/10.1145/ 1629890.1629907.
6. T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, WWW ’10, Association for Computing Machinery, Raleigh, North Carolina, USA, 2010, pp. 851–860, https://doi.org/10.1145/ 1772690.1772777, doi:10.1145/1772690.1772777.
7. Y. Tyshchuk, C. Hui, M. Grabowski, W.A. Wallace, Social media and warning response impacts in extreme events: results from a naturally occurring experiment, in: 2012 45th Hawaii International Conference on System Sciences, 2012, pp. 818–827, https://doi.org/10.1109/HICSS.2012.536.
8. S.E. Middleton, L. Middleton, S. Modafferi, Real-time crisis mapping of natural disasters using social media, IEEE Intell. Syst. 29 (2) (2014) 9–17, https://doi.org/10.1109/MIS.2013.126.
9. S. Muralidharan, L. Rasmussen, D. Patterson, J.-H. Shin, Hope for Haiti: an analysis of Facebook and Twitter usage during the earthquake relief efforts, Publ. Relat. Rev. 37 (2) (2011) 175–177, doi:10.1016/j.pubrev.2011.01.010, http://www.sciencedirect.com/science/article/pii/S0363811111000294.
10. M. Imran, F. Alam, U. Qazi, S. Peterson, F. Ofli, Rapid Damage Assessment Using Social Media Images by Combining Human and Machine Intelligence, arXiv preprint arXiv:2004.06675.
11. M.T. Niles, B.F. Emery, A.J. Reagan, P.S. Dodds, C.M. Danforth, Social media usage patterns during natural hazards, PloS One 14 (2) (2019), e0210484.
12. Internet Live Stats - Internet Usage & Social Media Statistics, 2014. https://www.internetlivestats.com/.
13. J.A. de Bruijn, H. de Moel, B. Jongman, M.C. de Ruiter, J. Wagemaker, J.C.J. H. Aerts, A global database of historic and real-time flood events based on social media, Sci. Data 6 (1) (2019) 1–12, doi:10.1038/s41597-019-0326-9, https://www.nature.com/articles/s41597-019-0326-9.
14. Y. Kryvasheyeu, H. Chen, N. Obradovich, E. Moro, P. Van Hentenryck, J. Fowler, M. Cebrian, Rapid assessment of disaster damage using social media activity, Sci. Adv. 2 (3) (2016), e1500779.
15. N. Pourebrahim, S. Sultana, J. Edwards, A. Gochanour, S. Mohanty, Understanding communication dynamics on Twitter during natural disasters: a case study of Hurricane Sandy, Int. J. Disaster Risk Reduct. 37 (2019) 101176, doi:10.1016/j. ijdrr.2019.101176, http://www.sciencedirect.com/science/article/pii/S2212420 918310434.

16. K. Leetaru, Visualizing Seven Years of Twitter's Evolution: 2012-2018, 2019. https://www.forbes.com/sites/kalevleetaru/2019/03/04/visualizing-seven-years- of-twitters-evolution-2012-2018/.
17. A. Gupta, H. Lamba, P. Kumaraguru, A. Joshi, Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy, in: Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 729–736.
18. F. Laylavi, A. Rajabifard, M. Kalantari, Event relatedness assessment of Twitter messages for emergency response, Inf. Process. Manag. 53 (1) (2017) 266–280, doi: 10.1016/j.ipm.2016.09.002, http://www.sciencedirect.com/science/article/pii/S0306457316303922.
19. N. Murzintcev, C. Cheng, Disaster hashtags in social media, ISPRS Int. J. Geo-Inf. 6 (7) (2017) 204.
20. F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, K. Tao, Semantics +filtering + search =twitcident. exploring information in social web streams, in: Proceedings of the 23rd ACM Conference on Hypertext and Social Media - HT '12, ACM Press, Milwaukee, Wisconsin, USA, 2012, p. 285, doi:10.1145/2309996.2310043, http://dl.acm.org/citation.cfm?doid=2309996.2310043.
21. X. Liu, B. Kar, C. Zhang, D.M. Cochran, Assessing relevance of tweets for risk communication, Int. J. Digital Earth 12 (7) (2019) 781–801, https://doi.org/10.1080/17538947.2018.1480670, doi:10.1080/17538947.2018.1480670.
22. H. Becker, M. Naaman, L. Gravano, Beyond Trending Topics: Real-World Event Identification on Twitter, ICWSM, 2011, https://doi.org/10.7916/D81V5NVX.
23. M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, P. Meier, Extracting Information Nuggets from Disaster- Related Messages in Social Media, 2013, p. 10.
24. K. Zahra, M. Imran, F.O. Ostermann, Automatic identification of eyewitness messages on twitter during disasters, Inf. Process. Manag. 57 (1) (2020) 102107.
25. A. Ilyas, Microfilters: harnessing twitter for disaster management, in: IEEE Global Humanitarian Technology Conference (GHTC 2014), IEEE, 2014, pp. 417–424.
26. M.-A. Kaufhold, M. Bayer, C. Reuter, Rapid relevance classification of social media posts in disasters and emergencies: a system and evaluation featuring active, incremental and online learning, Inf. Process. Manag. 57 (1) (2020) 102132.
27. M.A. Sit, C. Koylu, I. Demir, Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of hurricane irma, Int. J. Digital Earth 12 (11) (2019) 1205–1229.
28. M. Imran, F. Ofli, D. Caragea, A. Torralba, Using Ai and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions, 2020.
29. L. Spinsanti, F. Ostermann, Automated geographic context analysis for volunteered information, Appl. Geogr. 43 (2013) 36–44, doi:10.1016/j.apgeog.2013.05.005, http://www.sciencedirect.com/science/article/pii/S0143622813001185.
30. J.P. De Albuquerque, B. Herfort, A. Brenning, A. Zipf, A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management, Int. J. Geogr. Inf. Sci. 29 (4) (2015) 667–689.
31. J.A. de Bruijn, H. de Moel, A.H. Weerts, M.C. de Ruiter, E. Basar, D. Eilander, J. C. Aerts, Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network, Comput. Geosci. (2020) 104485.

32. P.M. Landwehr, K.M. Carley, Social media in disaster relief, in: Data Mining and Knowledge Discovery for Big Data, Springer, 2014, pp. 225–257.
33. J.P. Cangialosi, A.S. Latto, R. Berg, Hurricane Irma, Tech. Rep. AL112017, National Oceanic and Atmospheric Administration U.S. Department of Commerce, Jun. 2018. https://www.nhc.noaa.gov/data/tcr/AL112017_Irma.pdf.
34. L. Nguyen, Z. Yang, J. Li, G. Cao, F. Jin, Forecasting People's Needs in Hurricane Events from Social Network, IEEE Transactions on Big Data, 2019, 1–1ArXiv: 1811.04577. doi:10.1109/TBDATA.2019.2941887, http://arxiv.org/abs/1811 .04577.
35. U. S. N. H. Center, Costliest U.S. Tropical Cyclones Tables Update, Tech. Rep, National Oceanic and Atmospheric Administration, Jan. 2018. https://www.nhc. noaa.gov/news/UpdatedCostliest.pdf.
36. Introduction to Tweet JSON (2019). URL https://developer.twitter.com/en/ docs/tweets/data-dictionary/overview/intro-to-tweet-json.
37. Geo Objects (2019). URL https://developer.twitter.com/en/docs/tweets/data-di ctionary/overview/geo-objects.
38. M. Tomczak, Spatial interpolation and its uncertainty using automated anisotropic inverse distance weighting (idw)-cross-validation/jackknife approach, J. Geogr. Inf. Decis. Anal. 2 (2) (1998) 18–30.
39. X. Yang, X. Xie, D.L. Liu, F. Ji, L. Wang, Spatial interpolation of daily rainfall data for local climate impact assessment over greater sydney region, Adv. Meteorol. (2015).
40. R.J. Light, Measures of response agreement for qualitative data: some generalizations and alternatives, Psychol. Bull. 76 (5) (1971) 365.
41. M.L. McHugh, Interrater reliability: the kappa statistic, Biochemia medica, Biochem. Med. 22 (3) (2012) 276–282.
42. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556 [cs]ArXiv: 1409.1556. URL http://arxiv.org/ abs/1409.1556.
43. K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, arXiv:1512.03385 [cs]ArXiv: 1512.03385. URL http://arxiv.org/abs/1512.03385.
44. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, arXiv:1512.00567 [cs]ArXiv: 1512.00567. URL http://arxiv.org/abs/1512.00567.
45. A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C.J.C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., 2012, pp. 1097–1105. http://papers.nips.cc/paper/4824- imagenet-classification-with-deep-convolutional-neural-networks.pdf.
46. M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Columbus, OH, USA, 2014, pp. 1717–1724, https://doi.org/10.1109/CVPR.2014.222. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909618.
47. Twitter, About Verified Accounts, 2020. https://help.twitter.com/en/managing- your-account/about-twitter-verified-accounts. (Accessed 28 June 2020).
48. S. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica (Ljubljana) 31.

49. A. Liaw, M. Wiener, Classification and regression by randomForest, R. News 2 (3) (2002) 18–22.
50. J. Ye, J.-H. Chow, J. Chen, Z. Zheng, Stochastic gradient boosted distributed decision trees, in: Proceeding of the 18th ACM Conference on Information and Knowledge Management - CIKM ’09, ACM Press, Hong Kong, China, 2009, p. 2061, doi:10.1145/1645953.1646301, http://portal.acm.org/citation.cfm?doid=164595 3.1646301.
51. D.G. Kleinbaum, M. Klein, E.R. Pryor, Logistic Regression: a Self-Learning Text, third ed., Statistics in the health sciences, Springer, New York, 2010.
52. P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (10) (2012) 78, doi:10.1145/2347736.2347755, http://dl.acm.org/citation.cfm?doid=2347736.2347755.
53. J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (Feb) (2012) 281–305. http://www.jmlr.org/papers/v13/bergstr a12a.html.
54. D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods, in: 33rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Cambridge, Massachusetts, USA, 1995, pp. 189–196, doi:10.3115/981658.981684, https://www.aclweb.org /anthology/P95-1026.
55. A.D. Marco, R. Navigli, Clustering and diversifying web search results with graph- based word sense induction, Comput. Ling. 39 (3) (2013) 709–754, doi:10.1162/COLI_a_00148, https://www.aclweb.org/anthology/J13-3008.
56. S. Arora, Y. Li, Y. Liang, T. Ma, A. Risteski, Linear algebraic structure of word senses, with applications to polysemy, Trans. Assoc. Computat. Linguist. 6 (2018) 483–495, doi:10.1162/tacl_a_00034, https://www.mitpressjournals.org/doi/abs/1 0.1162/tacl_a_00034.
57. M. Imran, P. Mitra, C. Castillo, Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages, arXiv:1605.05894 [cs]ArXiv: 1605.05894. URL http://arxiv.org/abs/1605.05894.
58. C. De Boom, S. Van Canneyt, B. Dhoedt, Semantics-driven event clustering in Twitter feeds, in: Proceedings of the 5th Workshop on Making Sense of Microposts, vol. 1395, CEUR, 2015, pp. 2–9. http://hdl.handle.net/1854/LU-6887623.
59. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 [cs]ArXiv: 1301.3781. URL http://arxiv.org/abs/1301.3781.
60. Y. Goldberg, O. Levy, word2vec Explained: deriving Mikolov et al.’s negative- sampling word-embedding method, arXiv:1402.3722 [cs, stat]ArXiv: 1402.3722. URL http://arxiv.org/abs/1402.3722.
61. O. Ozdikis, P. Senkul, H. Oguztuzun, Semantic expansion of tweet contents for enhanced event detection in twitter, in: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012, pp. 20–24, https://doi.org/10.1109/ASONAM.2012.14, iSSN: null.
62. G. Salton, E.A. Fox, H. Wu, Extended boolean information retrieval, communications of the ACM. https://dl.acm.org/doi/abs/10.1145/182.358466.
63. D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, J. Weaver, Google street view: capturing the world at street level, Computer 43 (6) (2010) 32–38.

64. A.G. Rundle, M.D. Bader, C.A. Richards, K.M. Neckerman, J.O. Teitler, Using google street view to audit neighborhood environments, Am. J. Prev. Med. 40 (1) (2011) 94–100.
65. R. Lagerstrom, Y. Arzhaeva, P. Szul, O. Obst, R. Power, B. Robinson, T. Bednarz, Image Classification to Support Emergency Situation Awareness, Front. Robot. AI 3. doi:10.3389/frobt.2016.00054. URL https://www.frontiersin.org/articles/10.3389/frobt.2016.00054/full.
66. D.T. Nguyen, F. Ofli, M. Imran, P. Mitra, Damage assessment from social media imagery data during disasters, in: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM ’17, ACM Press, Sydney, Australia, 2017, pp. 569–576, doi:10.1145/ 3110025.3110109, http://dl.acm.org/citation.cfm?doid=3110025.3110109.
67. X. Li, H. Zhang, D. Caragea, M. Imran, Localizing and Quantifying Damage in Social Media Images, arXiv:1806.07378 [cs]ArXiv: 1806.07378. URL http://arxiv. org/abs/1806.07378.
68. A. Gupta, H. Lamba, P. Kumaraguru, A. Joshi, Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy, in: Proceedings of the 22nd International Conference on World Wide Web - WWW ’13 Companion, ACM Press, Rio de Janeiro, Brazil, 2013, pp. 729–736, doi:10.1145/2487788.2488033, http://dl.acm.org/citation.cfm?doid=2487788.2488033.
69. F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, Detection of GAN-generated fake images over social networks, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 384–389, https://doi.org/10.1109/MIPR.2018.00084, iSSN: null.
70. C. Buntain, J. Golbeck, Automatically identifying fake news in popular twitter threads, in: 2017 IEEE International Conference on Smart Cloud (SmartCloud), 2017, 208–215 ArXiv: 1705.01613. doi:10.1109/SmartCloud.2017.40, http://arxiv .org/abs/1705.01613.
71. X. Zhou, R. Zafarani, Fake News: A Survey of Research, Detection Methods, and Opportunities. URL https://arxiv.org/abs/1812.00315v1.
72. F. Masood, G. Ammad, A. Almogren, A. Abbas, H.A. Khattak, I. Ud Din, M. Guizani, M. Zuair, Spammer detection and fake user identification on social networks, IEEE Access 7 (2019) 68140–68152, https://doi.org/10.1109/ACCESS.2019.2918196.
73. M. Del Vicario, W. Quattrociocchi, A. Scala, F. Zollo, Polarization and Fake News: Early Warning of Potential Misinformation Targets, arXiv:1802.01400 [cs] ArXiv: 1802.01400. URL http://arxiv.org/abs/1802.01400.
74. A.H. Wang, Detecting spam bots in online social networking sites: a machine learning approach, in: D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M.Y. Vardi, G. Weikum, S. Foresti, S. Jajodia (Eds.), Data and Applications Security and Privacy XXIV, vol. 6166, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 335–342, doi:10.1007/978-3-642-13739-6_25, http://link.springer.com/10.1007/978-3-642-13739-6_25.
75. V.S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, F. Menczer, A. Stevens, A. Dekhtyar, S. Gao, T. Hogg, F. Kooti, Y. Liu, O. Varol, P. Shiralkar, V. Vydiswaran, Q. Mei, T. Hwang, The DARPA twitter

bot challenge, Computer 49 (6) (2016) 38–46, arXiv: 1601.05140. doi:10.1109/MC.2016.183, http://arxiv.org/abs/1601.05140.

76. P. Efthimion, S. Payne, N. Proferes, Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots, SMU Data Science Rev. 1(2). URL https://scholar.smu.edu/datasciencereview/vol1/iss2/5.
77. S.R. Sahoo, B.B. Gupta, Hybrid approach for detection of malicious profiles in twitter, Comput. Electr. Eng. 76 (2019) 65–81, doi:10.1016/j. compeleceng.2019.03.003, http://www.sciencedirect.com/science/article/pii/ S0045790618322766.
78. E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini, The rise of social bots, Communications of the ACM. URL https://dl.acm.org/doi/abs/10.1145/2818717.
79. A. Karami, V. Shah, R. Vaezi, A. Bansal, Twitter Speaks: A Case of National Disaster Situational Awareness, arXiv:1903.02706 [cs, stat]ArXiv: 1903.02706. URL http://arxiv.org/abs/1903.02706.
80. S. Mohanty, B. Biggers, S. Sayedahmed, E. Goldstein, R. Bunch, G. Chi, F. Sadri, T. McCoy, A. Cosby, Geolocated Tweets from florida, usa during Hurricane Irma, 2017 with relevance scores (Jan 2021). doi:10.6084/m9.figshare.11900325, https://figshare.com/articles/dataset/Geolocated_Tweets_from_Florida_USA_during_Hurricane_Irma_2017_with_Relevance_Scores/11900325/1.