Monitoring Misinformation on Twitter During Crisis Events: A Machine Learning Approach

Kyle Hunt, Puneet Agarwal, and Jun Zhuang*

Social media has been increasingly utilized to spread breaking news and risk communications during disasters of all magnitudes. Unfortunately, due to the unmoderated nature of social media platforms such as Twitter, rumors and misinformation are able to propagate widely. Given this, a surfeit of research has studied false rumor diffusion on Twitter, especially during natural disasters. Within this domain, studies have also focused on the misinformation control efforts from government organizations and other major agencies. A prodigious gap in research exists in studying the monitoring of misinformation on social media platforms in times of disasters and other crisis events. Such studies would offer organizations and agencies new tools and ideologies to monitor misinformation on platforms such as Twitter, and make informed decisions on whether or not to use their resources in order to debunk. In this work, we fill the research gap by developing a machine learning framework to predict the veracity of tweets that are spread during crisis events. The tweets are tracked based on the veracity of their content as either true, false, or neutral. We conduct four separate studies, and the results suggest that our framework is capable of tracking multiple cases of misinformation simultaneously, with F_1 scores exceeding 87%. In the case of tracking a single case of misinformation, our framework reaches an F_1 score of 83%. We collect and drive the algorithms with 15,952 misinformation-related tweets from the Boston Marathon bombing (2013), Manchester Arena bombing (2017), Hurricane Harvey (2017), Hurricane Irma (2017), and the Hawaii ballistic missile false alert (2018). This article provides novel insights on how to efficiently monitor misinformation that is spread during disasters.

KEY WORDS: Emergency management; fake news; machine learning; risk communications; social media

1. INTRODUCTION

With the rapid progression of Web 2.0 and its related innovations, billions of people around the world take to the online environment to retrieve news, share their opinions and emotions, follow along with their interests and hobbies, and even stream their favorite television shows. As a direct result of such technology, social media platforms have become increasingly popular due to their interactive two-way communications and advanced network dynamics. Within social media platforms, such as Twitter and Facebook, information is able to travel across the globe at extreme speeds, allowing users to obtain timely information related to politics, sports, economics, and all other imaginable topics. Some features of these platforms, such as "retweeting," "commenting," and "sharing" allow for information to spread throughout many different friend groups and social networks.

During crisis events of all magnitudes, including natural disasters, man-made disasters, terrorist

Department of Industrial and Systems Engineering, The State University of New York at Buffalo, Buffalo, NY, USA.

*Address correspondence to Jun Zhuang, Department of Industrial, and Systems Engineering, The State University of New York at Buffalo, Buffalo, NY 14260-2050, USA; jzhuang@buffalo.edu.

attacks, and disease spreading, it is critical that affected populations receive timely and valid communications. Given their efficient information diffusion features, social media platforms have been widely used for disseminating emergency and risk communications (Acar & Muraki, 2011; Houston et al., 2015; Mills, Chen, Lee, & Raghav Rao, 2009; Simon, Goldberg, & Adini, 2015; Vos et al., 2018). In times of crises, millions of people turn to social media for breaking news updates, evacuation planning, situational awareness, safety protocols, among many other emergency needs. Although there are many significant benefits associated with social media platforms, there are also certain characteristics which can lead to a dangerous social environment. Unfortunately, due to the unmoderated nature of Twitter and other social media platforms, misinformation and fake news often spread, reaching and influencing people around the world. During crisis events, when information integrity is of the upmost importance to the safety and well-being of affected citizens, misinformation often circulates throughout social media (Hunt, Agarwal, & Zhuang, 2019b; Hunt, Agarwal, Al Aziz, & Zhuang, 2020a). Such issues have catalyzed a significant amount of research to be carried out in this domain, with many works focusing on the dynamics of rumor spreading during disasters, and the related behaviors of social media users.

In this work, we define misinformation as any and all false information, including false rumors, false alerts, fake news, disinformation, and hoaxes (e.g., disinformation is false information which is intended to mislead the public, and is therefore a type of misinformation). To provide an example of misinformation that was spread on Twitter during a disaster, we turn to Hurricane Sandy, which was the deadliest and strongest Atlantic hurricane in 2012. During this immense storm which impacted the Bahamas, United States, Bermuda, and Canada, many false rumors were spread, including one which stated that the New York Stock Exchange building was flooded (Wang & Zhuang, 2018). Another false rumor during Hurricane Sandy stated that all bridges going into and out of Manhattan were being sealed off (DHS, 2018). Rumors such as these can cause unneeded chaos and confusion in time-sensitive situations.

Along with misinformation diffusion research, research on misinformation debunking and correction is gaining in popularity (Chua, Tee, Pang, & Lim, 2017; Hunt, Wang, & Zhuang, 2020b; Shin, Jian, Driscoll, & Bar, 2017; Takayasu et al., 2015; Wang & Zhuang, 2018). When misinformation is spread online and/or offline, major government, news, NGO, and emergency management agencies make statements and post to social media platforms in order to debunk the misinformation and provide the public with accurate content (Hunt et al., 2020b). For example, following the false rumors during Hurricane Sandy, the U.S. Federal Emergency Management Agency (FEMA) created a "Rumor Control" page on their website in order to dispel the inaccuracies and provide updated and thorough communications. FEMA used their Twitter account to disseminate this web page by posting 12 different tweets over the course of one week, with all of these tweets having a direct link to the Rumor Control page. In order to debunk misinformation, agencies must expend human resources and time in order to locate misinformation on social media, track the misinformation in order to understand its reach and impact, and formulate debunking messages. To the best of our knowledge, and through conversations with emergency managers, the current process of locating and choosing which misinformation to debunk is a manual process which requires an expert to sort through information and make decisions on what needs to be debunked.

With the spread of misinformation threatening the reliability of social media for emergency and risk communications, it is important to study new methods to debunk and control misinformation in an efficient manner. Many researchers have studied the classification, detection, and resolution of rumors on social media platforms, as summarized in Zubiaga, Aker, Bontcheva, Liakata, and Procter (2018). Although these works make significant contributions to the literature and knowledge related to rumors—which are defined as items "of circulating information whose veracity status is yet to be verified at the time of posting (Zubiaga et al., 2018)"—there is very limited literature which studies the detection (Jain, Sharma, & Kaushal, 2016) and termination of *misinformation* on social media platforms; especially during crisis events. Furthermore, to the best of our knowledge, there is no literature which studies the tracking of misinformation during crisis events. Given this gap in research, we are motivated to develop and analyze a system that can accomplish this. Such a system must be capable of predicting the veracity of tweets that are associated with a specific topic, which would in turn provide agencies and Twitter users with a tool that can support the monitoring and resolution of misinformation during perilous situations. By having information on the number and types of Twitter users who are spreading misinformation, agencies can make informed decisions on whether or not to release debunking information.

Two research questions will be answered in this study, including: RQ1: Can machine learning algorithms accurately predict the veracity of tweets which are related to misinformation? and RQ2: Can these algorithms be trained on a corpus of historical misinformation-related tweets in order to predict the veracity of tweets associated with a new case of misinformation? These research questions are motivated by the gaps in literature, and the need for more efficient methods to monitor, debunk, and study misinformation.

In order to answer these questions, we develop, test, and analyze a machine learning pipeline that is capable of tracking misinformation based on a given topic. We first collected Twitter data for six cases where misinformation was spread during crisis events, including two false rumors that were spread during hurricanes, three false rumors that were spread during terrorist attacks, and one false alert that was spread by a governmental agency. The data were then processed and labeled, and subsequently used to train machine learning algorithms. By accurately predicting the veracity of tweets, the machine learning pipeline allows analysts to track the spreading and debunking of misinformation in order to make informed decisions regarding control efforts and further communications. This study fills the research gap by offering a novel approach to track misinformation on Twitter during crisis events. Although the results of this research will be particularly beneficial to agencies involved in the dissemination of risk and emergency communications, they will also be applicable to researchers who wish to automatically annotate social media data in order to conduct big data studies.

The rest of the article is organized as follows: Section 2 introduces the related literature and exposes the gap that this study bridges; Section 3 presents the research methodology, including the background of the misinformation cases, the data collection and processing, and the machine learning pipeline; Section 4 presents the results and analysis; and Section 5 provides a concluding discussion and future research directions.

2. LITERATURE REVIEW

2.1. Social Media as an Emergency Response System

Social media platforms, such as Twitter and Facebook, have evolved as critical knowledge resources for emergency responders, decisionmakers, and the general public during crisis events. These platforms allow citizens to engage in the process of emergency management by enabling them to easily disseminate real-time information to the public, and also enabling them to access information from within these platforms (Simon et al., 2015). During disasters, social media provides a cost-effective mechanism to collect time-sensitive data by allowing information gathering to be crowd-sourced (Maresh-Fuehrer & Smith, 2016). Social media technologies have been successfully leveraged by responding organizations to establish coordination with various external aid agencies for the effective acquisition, use, sharing, and maintenance of knowledge (Yates & Paquette, 2010). Social media outperforms the traditional mass media given its efficient and timely information diffusion features, and interactive two-way communications (Fraustino, Liu, & Jin, 2012). As a result of these benefits, extensive research has investigated the use of social media during disasters (Abedin, Babar, & Abbasi, 2014; Houston et al., 2015; Lundgren & McMakin, 2018; Wang & Zhuang, 2017), including for crisis communications (Bruns & Burgess, 2014; Lin, Margolin, & Wen, 2017), information credibility (Gupta & Kumaraguru, 2012; Spence, Lachlan, Lin, & del Greco, 2015), and situational awareness (Vieweg, Hughes, Starbird, & Palen, 2010; Yuan, Guan, Huh, & Lee, 2013).

2.2. Rumors and Misinformation on Social Media

2.2.1. Propagation Dynamics

Numerous studies have characterized the emergence and propagation of rumors in social media platforms. Maddock et al. (2015a) used the concept of "multidimensional signatures" to characterize different types of rumors. These signatures include several quantitative measures including temporal progression of rumor-related behaviors, URL domain diversity, domain propagation over time, lexical diversity, and geolocation features. Liao and Shi (2013) explored the dynamics of rumor transmission in China's largest microblogging system, Sina Weibo, and identified four major categories that describe how users intervene in rumor discussions: providing information, expressing emotions, sharing opinions, and analyzing and interpreting situations. Zubiaga, Liakata, Procter, Hoi, and Tolmie (2016b) analyzed a data set of 330 rumor threads associated with nine newsworthy events to understand the role of different types of users in the spreading and debunking processes throughout the life cycle of a rumor Arif et al. (2016) applied a mixed-methods framework based on volume, exposure, and content production to study the dynamics of rumors. This study demonstrates a holistic understanding of rumor transmission by investigating the relationship between rumor content and the people engaging with the rumor content. Cheng, Liu, Shen, and Yuan (2013) found that the diffusion of rumors in online social networks is a function of the strength of ties between users, where the possibility of a rumor spreading is more likely across strong ties in a network. Studies conducted by Oh, Agrawal, and Rao (2013) on rumor mongering show that the effect of source ambiguity (the lack of an official source) on rumoring is much more significant than that of content ambiguity (lack of persuasive statements in Twitter posts), and anxiety. Vosoughi, Roy, and Aral (2018) analyzed the diffusion dynamics of true and false rumors and found that false rumors (misinformation) propagated significantly faster and deeper as compared to true rumors in all categories of information; namely, political news, terrorism, natural disasters, science, urban legends, entertainment, and financial information. Due to the surfeit of recent research that has been conducted in rumor and misinformation diffusion, it is of critical importance to utilize past research in the development of misinformation management tools. Such tools would be of great value to industry and academic researchers, as well as government agencies in their efforts to make informed decisions regarding the control of misinformation within social media.

2.2.2. *Spread of Rumors and Misinformation during Crisis Events*

Due to the vast use of social media for emergency communications, many researchers have studied the propagation of false information and rumors on social media platforms (Castillo, Mendoza, & Poblete, 2013; Procter, Vis, & Voss, 2013; Starbird, Maddock, Orand, Achterman, & Mason, 2014; Li & Sakamoto, 2015; Zubiaga et al., 2016b). Following the 2011 Great East Japan Earthquake, misinformation propagated throughout Twitter stating that rain in the earthquake's aftermath may include harmful chemicals. Due to this, there was a warning which recommended using umbrellas outdoors (Tanaka, Sakamoto, & Matsuka, 2012). Researchers located a rumor correction tweet which aided in the termination of the false information, and they were able to create a model to estimate the rumor infection rate, along with the number of people who still believed in the false rumor after the correction tweet was posted (Takayasu et al., 2015).

During Hurricane Sandy and the Boston Marathon bombing (2013), Twitter users performed poorly in detecting false rumors and rushed to spread the rumored news, further contaminating Twitter's network (Wang & Zhuang, 2018). These results were supported by Starbird et al. (2014), where the authors studied three false rumors that spread during the Boston Marathon bombing. The results from this study suggested that Twitter users did not do well in distinguishing truth from hoax, and that the propagation of the misinformation overpowered the debunking efforts. Following the 2010 Chile earthquake, tweets related to news topics were compared to tweets related to rumors, and the propagation patterns were found to be much different. The authors concluded that this was because rumors are questioned more than news (Mendoza, Poblete, & Castillo, 2010).

Many research works which study rumor and misinformation diffusion take advantage of content analysis and manual coding schemes to classify and analyze their data sets (Hunt et al., 2020b; Oh et al., 2013; Starbird et al., 2014; Wang & Zhuang, 2018; Zubiaga et al., 2018), and such research has continued to grow in recent years.

2.3. Approaches to Automated Rumor and Misinformation Classification

Once a case of misinformation is identified within social media networks, the subsequent online communications associated with the misinformation have to be monitored in order to take timely actions and contain the spread of the misinformation. These communications can help the emergency response and government organizations quantify the extent of misinformation spread based on the number of users who are offering false or valid information. Hence, the problem of automated misinformation tracking can be conceptualized as a classification task that consists of determining the veracity of the information that an individual user posts with respect to the detected case of misinformation.

Over the last few years, the classification task in the context of social media has attracted many studies (Hunt, Agarwal, & Zhuang, 2019a). Qazvinian, Rosengren, Radev, and Mei (2011) used Bayesian classifiers to classify the rumor-related tweets as either confirming, denying, or doubtful. In this study, the authors used different features categorized as "content," "network," and "Twitter-specific memes" to train the algorithms. It was found that the content-based features gave a better performance as compared to the other features. Hamidian and Diab (2015) implemented decision trees on the data set created by Qazvinian et al. (2011) to perform the rumor classification task. In addition to the features adopted by Qazvinian et al. (2011), Hamidian and Diab (2015) introduced pragmatic attributes such as entities, events, sentiments, and emoticons to achieve higher performance scores. Zeng, Starbird, and Spiro (2016) used three different machine learning models—Logistic Regression, Gaussian Naive Bayes, and Random Forest—for automatically classifying rumor-related tweets as one of three mutually exclusive categories: affirm, deny, or neutral. The best performance scores were achieved using Random Forests. Zubiaga et al. (2016a) used two different sequential classifiers: a linear-chain conditional random fields (CRF), and a tree CRF to address stance classification of rumor-related tweets within conversational threads. Their findings show that sequential approaches perform substantially better as compared to nonsequential baselines such as Support Vector Machines, Random Forests, Naive Bayes, and Maximum Entropy classifiers. Kochkina, Liakata, and Augenstein (2017) proposed a long short-term memory (LSTM)-based sequential model to exploit the conversational structure of social media threads for stance classification. Kwon, Cha, Jung, Chen, and Wang, (2013) built three classifiers based on Decision Trees, Random Forests, and Support Vector Machine to determine whether a topic is a rumor or not. Their models were based on the temporal, structural, and linguistic features, and were able to achieve higher performance scores as compared to the other state-of-the-art works on rumor classification. Zhang et al. (2015b) focused exclusively on health rumors for rumor veracity classification task. This study investigated the correlation between features and veracity of rumors based on logistic regression. Their findings indicate that a rumor is more likely to be true if it contains elements such as numbers, source cues, and hyperlinks, while the presence of images within the posts is negatively correlated with true rumors.

As evident from the scientific literature, stance and veracity classification of rumors have been well studied by numerous researchers. The veracity classification task aims to determine the actual truth value of the rumor, while the stance classification task aims to determine the type of orientation expressed by a user with respect to the rumor's veracity. To the best of our knowledge, no prior studies have provided an automated framework to determine the veracity of tweets which are related to misinformation. By monitoring the veracity of misinformation-related tweets, it becomes clear if users are continuing to post false information, or if users are beginning to correct the false information. Such an automated monitoring framework allows emergency response organizations and concerned authorities to make timely decisions by analyzing the posts. The existing studies devoted to the classification of social media data are mainly experimented on data sets that are related to political news, entertainment, terrorism, urban legends, and financial information (Hamidian & Diab, 2015; Qazvinian et al., 2011; Vosoughi et al., 2018). This is primarily due to the lack of data sets related to the spread of misinformation during crisis events such as natural disasters, man-made disasters, and false alerts. In such events, the necessity of an automated misinformation monitoring framework can be deemed as absolutely necessary to prevent widespread panic and/or confusion among the people. Aside from the literature on rumor and misinformation propagation, we also draw motivation from the extensive literature on risk communications in order to understand emotional factors (Xie, Wang, Zhang, Li, & Yu, 2011), the effect of such communications (Fischhoff, Gonzalez, Small, & Lerner, 2003; Lazo, Bostrom, Morss, Demuth, & Lazrus, 2015), as well as trust issues in emergency communications (Murayama, Saito, & Nishioka, 2013). Such literature and knowledge is critical in the development of a technology-based intervention system for crisis-related misinformation.

In this article, we address the existing gaps in the literature by providing a framework for tracking misinformation that is spread during crisis events. Our framework is extensively tested on large-scale data sets which we collected from six different cases of misinformation that were spread during crisis situations. In order to validate the novelty and robustness of the framework, we conduct four separate studies.

In the first study, we combine all of the data sets to select a machine learning model that performs the best in predicting the veracity of tweets. In the second study, we use different partitions for the testing and training data in order to understand how sensitive the model is to sparse training data. For example, we first use 90% of the data for training and 10% for testing, and then vary the partitions until we finally train with 10% of the data and test with the remaining 90%. In the third study, we train and test the model on all of the six cases separately in order to analyze how the model performs given a single topic. In the final study, we use data from five cases plus a percentage of the sixth case to train the model, and we then test the model on the remainder of the data from the sixth case. This study aims to identify if including features from the five cases in training can help to predict the sixth case. The results and conclusions of this research address the practical applicability of the developed framework for the surveillance and overall integrity of emergency and risk-related communications. This framework is also directly adoptable by social media researchers who wish to automate the process of annotating their text-based data sets.

3. RESEARCH METHODOLOGY

3.1. Background of Selected Misinformation Cases

To ensure the robustness and feasibility of the machine learning framework presented in this research, we study six cases of misinformation that spread during crisis events, including five false rumors and one false alert. These crisis events consist of the Boston Marathon bombing (2013), Manchester Arena bombing (2017), Hurricane Harvey (2017), Hurricane Irma (2017), and the Hawaii ballistic missile false alert (2018). The criteria for choosing and collecting these data sets was based upon their large-scale news coverage and the availability of the data on Twitter. The two false rumors from the Boston Marathon bombing were broadcast across the online environment, and were identified through news outlets and social media platforms (Sager, 2013). The Manchester Arena bombing false rumor was identified through major news outlets, such as The New York Times and CBS (Qiu, 2017). For the Hurricane Harvey and Hurricane Irma false rumors, the cases were identified on FEMA's Rumor Control pages (FEMA, 2017a, 2017b). News from the 2018 Hawaiian incoming missile false alert was broadcasted online, on the radio, and on television. Due to the fact that data collection for this study began in 2013, no major crisis events before this year are considered. Similarly, this study does not consider an exhaustive list of all cases where misinformation spread during crisis events between the years of 2013 and 2019. Thorough descriptions of the selected events and related misinformation follows:

Boston Marathon Bombing Sandy Hook Rumor. On April 15, 2013, the United States was struck by an act of terrorism. Two homemade pressure cooker bombs were detonated near the finish line of the Boston Marathon, killing three people, and significantly injuring many more (including 16 people who lost limbs). During the chaos that ensued, many false rumors were spread. One the most prominent false rumors stated that an eight year-old girl was killed in the bombings while she was running in remembrance of the 2012 Sandy Hook School shooting victims.

Boston Marathon Bombing Donation Rumor. Following the 2013 Boston Marathon bombing, one false rumor directly utilized Twitter. A fake account named @_BostonMarathon was created and posted a tweet which read "For every retweet we receive we will donate $1.00 to the #BostonMarathon victims." Many users ended up retweeting the post believing that it would aid in the recovery from the disaster, but in fact this account had no intentions to donate any money. Twitter eventually suspended the fraudulent account, and warnings spread throughout Twitter to look out for other accounts like this one and their related dissemination of misinformation.

Manchester Bombing Holiday Inn Rumor. On May 22, 2017, Ariana Grande was performing a concert in the Manchester Arena in England. When the concert concluded and attendees were beginning to exit the venue, a suicide bomber detonated explosives that were attached to his body. The bombing led to 23 deaths and over 139 wounded people, and was the deadliest terrorist attack in England since the 2005 London bombings. After the bombing, a false rumor was spread on Twitter and Facebook which stated that unaccompanied children were being taken to safety at a local Holiday Inn. To debunk this misinformation, a Holiday Inn representative had to make a public statement, clearly informing the public that there were no unaccompanied children at the hotel.

Hurricane Harvey Immigration Rumor. On August 25, 2017, Hurricane Harvey made landfall in Texas, United States. Before and during Hurricane Harvey, there was legislation due to be passed in

Texas that was aiming to increase anti-immigration policies. As a result, some people began to inquire about identification checks at evacuation shelters, and a false rumor developed throughout social media and Texas which stated that shelters were going to be checking IDs. This rumor proved to be very threatening, as many undocumented immigrants were fearful of going to shelters due to their lack of citizenship and the potential threat of deportation. Many government and news accounts, including Houston's official Twitter account, posted rumor debunking tweets in order to help contain the spread of the misinformation.

Hurricane Irma Immigration Rumor. Immediately succeeding Hurricane Harvey, Hurricane Irma was generating immense damage across the Caribbean on its path toward the United States. On September 10, 2017, Hurricane Irma made landfall in Cudjoe Key, Florida, bringing along deadly storm conditions. On September 6, 2017, a Polk County Sheriff posted on Twitter saying that he would be checking identifications at all evacuation centers in his jurisdiction, which caused anger and fear among citizens and undocumented immigrants. Although the Sheriff's tweets were factual in their content, and he did not spread any misinformation, many citizens and undocumented immigrants inferred that he was checking IDs to primarily scare undocumented immigrants from seeking safety in Polk County shelters. As a result, a false rumor began to contaminate Twitter which stated that undocumented immigrants were not allowed in shelters. The Sheriff later posted a new tweet to clarify his intentions and let the public know that he was not targeting undocumented immigrants with his initial message. Besides this clarification tweet, many debunking efforts were made in order to help comfort the population. The City of Miami, FEMA, the United States Department of Homeland Security (Customs and Border Protection as well as Immigration and Customs Enforcement), and other organizations posted information to let the undocumented population know that it was safe for them to seek shelter.

Hawaii Missile False Alert On January 13, 2018, Hawaii's Emergency Management Agency sent out an emergency alert to cell phones, televisions, and radio stations stating that a ballistic missile was headed toward the islands. The emergency notification that was sent to cell phones across Hawaii read "BALLISTIC MISSILE THREAT INBOUND TO HAWAII. SEEK IMMEDIATE SHELTER. THIS IS NOT A DRILL." A second alert that was sent 38 minutes later informed the public that the first notification was false, and that there was no incoming missile. The notification that a ballistic missile was headed toward Hawaii was false information upon its release, and there was no expert or informed personnel who authorized the dissemination of this information. The false alert was caused by human error; an employee of the Hawaii Emergency Management Agency pressed the wrong button during a routine preparatory drill, and was later fired for this mistake (Park, Allen, Davidson, & Turrell, 2018). Hawaii governor David Ige and other government authorities posted tweets to clarify the misinformation.

3.2. Data Collection

Twitter's Search API (https://developer.twitter.com/en/docs) and Python were used for collecting all of the tweets in this research. Twitter's Standard Search API returns tweets from the previous seven days based on user-specified search criteria (queries). The API does not return an exhaustive list of tweets; therefore our data contain a sample of the tweets related to our queries. To counter this problem and collect more complete data sets, the collection took place over a 28-day window for every case, with collection done every three days using the same search criteria every time. This method allows for data to be collected at least twice for every day in the 28-day window (excluding the last three days before collection ended). Although this method still does not supply every related tweet and is computationally expensive, it gives us less limited data sets, and allows us to capture many tweets which may have been deleted after any of our given collections (Maddock, Starbird, & Mason, 2015b). The different search periods for the six cases, along with the number of tweets collected for each case, are provided in Table I. In total between the six cases, we collected 15,952 misinformation-related tweets (after removing unrelated tweets from the data, which is explained in Section 3.3). The queries used for all cases were a combination of case-insensitive keywords and hashtags (e.g., immigration and #harvey; hawaii and #missile). The queries were chosen following an extensive Twitter Advanced Search (https://twitter.com/search-advanced?lang=en) to find the major keywords and hashtags that identified tweets related to the false rumors and false alert. The exact queries used for all of the cases were searched in English, and Table II provides examples of tweets that were collected from all of the events.

Table I. Collection Dates and Total Tweets Collected for the Six Events

Crisis Event	Collection Began	Collection Ended	Collected Tweets
Boston Marathon Bombing Sandy Hook Rumor	April 18, 2013	May 15, 2013	2,101
Boston Marathon Bombing Donation Rumor	April 18, 2013	May 15, 2013	650
Manchester Bombing Holiday Inn Rumor	May 25, 2017	June 21, 2017	3,882
Hurricane Harvey Immigration Rumor	August 28, 2017	September 24, 2017	2,034
Hurricane Irma Immigration Rumor	September 9, 2017	October 6, 2017	594
Hawaii Missile False Alert	January 14, 2018	February 10, 2018	6,691
Total collected tweets			**15,952**

3.3. Data Coding Scheme

Utilizing latent content analysis and following the rules suggested by Hunt et al. (2020b) and Krippendorff (2013), the text of each tweet was coded to classify the information within it. Three researchers (“coder 1,” “coder 2,” and “coder 3”) participated in the coding process for all of the tweets. The coders were required to become familiar with all six of the misinformation cases in this study before coding began. Coders 1 and 2 independently coded all of the tweets into the following four mutually exclusive classes (relating to the veracity of the tweets’ content): true, false, neutral, and nonrelated. A rubric for the coding scheme is provided in Table III.

After the independent coding was completed, coder 3 then cross-validated all of the tweets in which coders 1 and 2 disagreed on the class (Dutta, Kwon, & Rao, 2018). There was no disagreement between tweets that were classified as “nonrelated,” and therefore these tweets were immediately discarded as they had no relevance to this research, and were purely noise from the data collection. For the 15,952 related tweets that were coded as “true,” “false,” or “neutral,” there ended up being 3,391 instances of disagreement between coders (78.7% intercoder agreement; Cohen’s kappa = 0.67 (Dutta et al., 2018)), and coder 3 determined which class was more prominent in the content of these tweets. The final results from coding are provided in Table IV, and the total amount of human time required in the coding process was 54 hours (22 hours for coder 1, 20.5 hours for coder 2, and 11.5 hours for coder 3).

3.4. Learning Twitter Data

For all machine learning tasks, Python programming language was used along with popular packages such as NumPy, pandas, Natural Language Toolkit (NLTK), and scikit-learn. Jupyter Notebook was used for the development and hosting of the machine learning pipeline. All of the machine learning tasks were carried out on a Dell XPS Desktop running on Windows 10 Enterprise (64-bit), containing 32 GB of RAM, an i7-4770 CPU, and 2 TB of hard drive.

3.4.1. Data Preprocessing

The textual content of the tweets was preprocessed to ensure that the extracted information was free from noise and inconsistency. The elements that were irrelevant for determining the nature of the tweets were removed/transformed so that the weight carried by them in comparison to the textual information was minimized. These elements consist of twitter handles, tags for trending topics, multiple white spaces, punctuation marks, numbers, special characters, upper case characters, URLs, and stop words. Twitter users are acknowledged on Twitter using twitter handles, represented by @user. The topics containing popular hashtags are called trending topics, represented by #topic. In these cases, special characters @ and # do not provide any useful information regarding the user’s orientation in an event of misinformation. Due to this, twitter handles are transformed to “AT_USER” and trending topics are transformed to “topic.” Elements such as multiple white spaces, punctuation, special characters and numbers act as noise in the text data and hence, are completely removed. Upper case characters are transformed into lower case characters to maintain data consistency. All URLs are converted to string “URL” to remove irrelevant information. From the tweet corpus, stop words (frequently occurring words

Table II. Examples of Misinformation-Related Tweets from All of the Crisis Events

Data Set	Example Tweets
Manchester Bombing Holiday Inn Rumor	1. Holiday Inn reports there were no unaccompanied children taken to the hotel last night. Who would report such a lie? #Manchester 2. Holiday Inn near Manchester Arena have taken 50+ kids who have been separated from their guardians tonight #ManchesterArena 3. I believe so, so many children too. Holiday Inn have taken in 50+ kids without guardians too.
Hurricane Harvey Immigration Rumor	1. We will not ask for immigration status or papers from anyone at any shelter. This rumor is FALSE! 2. Very important, anyone at any shelter will ask for immigration status or paper @KHOU 3. ICE is using Hurricane Harvey to set up immigration check points at disaster shelters to detain people who don't have documents
Hurricane Irma Immigration Rumor	1. Undocumented immigrants are too afraid to seek shelter from Hurricane Irma 2. When it comes to rescuing people in the wake of Hurricane Irma, immigration status is not and will not be a factor. 3. Even Hurricane Irma was about race. Immigrants were going to be locked up if they sought shelter. Their choice was jail or live.
Hawaii Missile False Alert	1. Hawaii missile threat was a "mistake" according to officials 2. GUYS HAWAII JUST GOT AN ALERT FOR A BALLISTIC MISSILE INBOUND!! THIS ISNT A DRILL! 3. Really hope that missile threat to Hawaii is a hoax
Boston Marathon Bombing Sandy Hook Rumor	1. R.I.P. to the eight-year-old girl who died in Boston's explosions, while running for the Sandy Hook kids 2. The "girl" that died while running for sandy hook victims is a boy. Just goes to show how incorrect the social media can display things. 3. All of you spreading photos of the boy and girl running saying they're victims of the Boston bombing, did you even confirm it was them?
Boston Marathon Bombing Donation Rumor	1. RT @_BostonMarathon: For every retweet we receive we will donate $1.00 to the #BostonMarathon victims 2. Fake Boston Marathon account going around saying it will donate $1 to the victims for each RT. Hate this stuff. 3. Beware of a FAKE Boston Marathon account going around saying it will donate $1 to the victims for each RT they get.

that add little or no value in prediction, e.g., "the," "an," and "and"), are also removed.

3.4.2. Feature Extraction

The next step after cleaning text data was to encode words in tweets in the form of numerical features. In this article, *n*-gram model (Cavnar & Trenkle, 1994) and term frequency-inverse document frequency (tf-idf) (Song, Guo, Hunt, & Zhuang, 2020) are used to construct textual features. In computational linguistics, an *n*-gram is a sequence of *n* words in a given sample of text. This model helps to boost predictive performance by taking into account the sequence of words, in contrast to just using singular words. Tf-idf weighting helps to quantify the importance of *n*-grams that appear in a given tweet corpus. Equation (1) gives the classical formula of tf-idf used in weighting *n*-grams.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right), \tag{1}$$

where $w_{i,j}$ is the weight for *n*-gram i in tweet j, N is the number of tweets in the corpus, $tf_{i,j}$ is the term frequency of *n*-gram i in tweet j, and df_i is the tweet frequency of *n*-gram i in the tweet corpus.

Table III. Coding Rubric Used to Define the Different Categories

Class	Definition
True	Contains valid content regarding the misinformation topic. These tweets debunk misinformation by delivering correct information to Twitter's network.
False	Contains false content regarding the misinformation topic. These tweets spread misinformation by delivering incorrect information to Twitter's network.
Neutral	Contains comments, questions, or shares opinions on the topic. These tweets **do not** offer true or false information.
Nonrelated	Not related to the misinformation topic in any way. These tweets are removed from the data set.

Table IV. Results from Data Coding

Data Set	True	False	Neutral
Boston Marathon Bombing Sandy Hook Rumor	858	994	249
Boston Marathon Bombing Donation Rumor	262	233	155
Manchester Bombing Holiday Inn Rumor	299	3,498	85
Hurricane Harvey Immigration Rumor	1,474	99	461
Hurricane Irma Immigration Rumor	263	80	251
Hawaii Missile False Alert	4,269	471	1,951
Total tweets per class	7,425	5,375	3,152

The number of times that an n-gram occurs in a tweet is called its term frequency (tf). Inverse document frequency (idf) is a measure of how much information the given n-gram conveys, that is, if it is common or rare across all tweets. So, an n-gram that is prevalent throughout the entire tweet corpus will have a lesser tf-idf value. This approach allows for the generation of a weighted numerical representation of text-based data, that also takes into account sequences of words.

3.4.3. *Feature Selection*

The next step is the feature selection process where the n-grams that contribute most in the prediction of the output variable are selected. As a result of this process, irrelevant and redundant n-grams are removed, thereby leading to a better performance in terms of predictive accuracy and speed of learning (Guyon & Elisseeff, 2003; Liu & Motoda, 2012). While developing the machine learning pipelines, the candidates selected for extracting textual features were unigrams, bigrams, and trigrams. Among these candidates, the prediction performance achieved using bigrams is found to be the best for each crisis event. The top ranked bigrams by each event case are shown in Table V.

In this article, the chi-squared (χ^2) (Yang & Pedersen, 1997) statistic is used to assess the importance of each n-gram on the output variable. The χ^2 statistic can be used to test whether the occurrence of a specific n-gram and the occurrence of a specific class in the output variable are independent. Given a tweet T, the χ^2 statistic is defined as follows:

$$\chi^2(T, j, c) = \sum_{e_j \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_j e_c} - E_{e_j e_c})^2}{E_{e_j e_c}}, \qquad (2)$$

where N is the observed frequency in T, E is the expected frequency, e_t takes the value of 1 if the tweet contains n-gram j and 0 otherwise, and e_c takes the value of 1 if the tweet corresponds to class c and 0 otherwise. The χ^2 statistic is calculated for each n-gram in the training corpus, and the top ranked n-grams that have the best χ^2 scores are selected to train the models.

In addition to the text-based features, user-specific features, namely, the number of likes and retweets that a tweet gets, the number of followers that a user has on Twitter, and the status of the user on Twitter (verified/unverified), are used to train the algorithms. Number of likes, retweets, and followers are in the form of numerical variables, while the status of the user takes the form of a categorical variable with a value of 1 for verified status and a value of 0 for unverified status. Table VI reports summary statistics

Table V. Top Ranked Bigrams for the Experimental Data Sets

Crisis Event	Top Ranked Bigrams
Boston Marathon Bombing Sandy Hook Rumor	year old, sandy hook, running sandy, girl dies, died boston, boston explosions, girl running, little girl, kids prayforboston, died running
Boston Marathon Bombing Donation Rumor	at_user retweet, retweet receive, receive donate, bostonmarathon victims, donated victims, gets donated, fake boston, marathon account, donate 00, marathon bombing
Manchester Bombing Holiday Inn Rumor	holiday inn, inn manchester, children guardians, 60 children, taking children, inn missing, separated parents, manchester taking, children holiday, children separated
Hurricane Harvey Immigration Rumor	immigration status, undocumented immigrants, ask immigration, seek shelter, shelter harvey, status papers, immigrants arrested, city houston, papers shelter, hurricane harvey
Hurricane Irma Immigration Rumor	hurricane irma, undocumented immigrants, seek shelter, afraid seek, immigrants afraid, shelter hurricane, immigration status, immigration enforcement, enforcement areas, miami dade
Hawaii Missile False Alert	missile alert, ballistic missile, hawaii missile, false alarm, missile threat, false missile, hawaii false, missile warning, alert hawaii, incoming missile

Table VI. Summary Statistics for the Experimental Data Sets Based on User-Specific Features

Crisis Event	# Likes per Unique Tweet	# Retweets per Unique Tweet	# Followers per Unique User Account	% Verified Users
Boston Marathon Bombing Sandy Hook Rumor	2	7	3,402	1%
Boston Marathon Bombing Donation Rumor	1	11	7,889	5%
Manchester Bombing Holiday Inn Rumor	37	50	9,292	3%
Hurricane Harvey Immigration Rumor	210	102	67,729	13%
Hurricane Irma Immigration Rumor	26	16	14,446	6%
Hawaii Missile False Alert	17	7	52,708	8%

for the experimental data sets based on user-specific features. In this table, it is observed that the data sets with a higher percentage of verified users tend to have a higher average number of followers per unique user account.

3.4.4. Machine Learning Algorithms

After all of the relevant features were extracted and selected, seven machine learning algorithms were trained and tested on the collected data; namely, *k*-nearest neighbors (kNN), decision tree (DT), random forest (RF), XGBoost (XGB), AdaBoost (AB), support vector machine (SVM), and multilayer perceptron (MLP). The purpose of this was to conduct a comparative study of their performances in terms of predictive power and speed of training. Such an analysis was used to identify the best performing algorithm in context to tracking misinformation. These algorithms were chosen to cover a wide array of concepts/techniques that are used in predictive analytics, namely, bagging, boosting, maximum-margin hyperplane, neural networks, and nearest-neighbors heuristics.

3.4.5. Cross-Validation and Hyperparameter Tuning

In this article, *k*-fold cross-validation (Agarwal, Tang, Narayanan, & Zhuang, 2020) is used on the training data as a resampling procedure to estimate the skills of the models on unseen test data. It is a popular method because it results in a less biased and less optimistic estimate of the models skill (Jung, 2018). This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal sizes. The first fold is treated as a validation set, and the method is fit on the remaining $k - 1$ folds. The number of models to be estimated increases as k increases, while a smaller k results in a smaller training set, which may pose a problem if the size of available data is small (Bergmeir, Costantini, & Benítez, 2014). Typical choices of k are 5 or 10 (Friedman, Hastie, & Tibshirani, 2001). In this study, the value of k is chosen to be 5 so that it results in a model skill estimate with low bias and moderate variance. Furthermore, the splitting of data into folds is done by ensuring that each fold has the same proportion of observations with a given class outcome value. This strategy is called stratified cross-validation.

A hyperparameter is an external characteristic of a model whose value cannot be estimated from data during the learning process. The values of the hyperparameters are set before the learning process of a model begins. In this article, grid-search is used to find the optimal combination of model hyperparameters which will result in highly accurate predictions.

3.4.6. Performance Metrics

In this article, we use three standard performance indicators, namely, precision (positive predictive value), recall (sensitivity), and F_1 score to evaluate the performance of the algorithms. Let TP and TN stand for true positive and true negative, respectively, and FP and FN stand for false positive and false negative, respectively. Using these values, we can compute precision (p) and recall (r) as follows:

$$p = \frac{TP}{TP + FP}, \tag{3}$$

$$r = \frac{TP}{TP + FN}. \tag{4}$$

F_1 score is a well-established classification performance measure that conveys a balance between precision and recall (Zhang, Wang, & Zhao, 2015a). It is known to be a more informative and transparent metric as compared to classification accuracy in problems that exhibit a class imbalance. When only one class is considered, the standard F_1 score is defined as the harmonic mean of precision and recall as shown in Equation (5).

$$F_1 \text{ Score} = \frac{2pr}{p + r}. \tag{5}$$

As multilabel classification can be decomposed into distinct binary classification problems, the F_1 score, as defined in Equation (5), can also be calculated separately for each class. A macro-averaged F_1 score is achieved simply by averaging the scores over the classes, that is, if m is the number of classes and F_{1_i} is the F_1 score for the class $i \in \{1, \ldots, m\}$,

$$\text{Macro average } F_1 \text{ Score} = \frac{\sum_{i=1}^{m} F_{1_i}}{m}. \tag{6}$$

In this research, the labels to be classified have an unequal distribution, as provided in Table IV. To account for a fairer evaluation of the algorithms, we use macro-averaged precision, recall, and F_1 scores. The performance metrics treat all classes equally regardless of the number of records within a class. A macro-average scaling of scores will help to select the algorithm that performs the best across all of the different labels (Zubiaga et al., 2018).

A general schematic of the framework that is developed in Section 3 and analyzed throughout this study is provided in Fig. 1. This framework is implemented using pipelines to automate the machine learning workflows. These pipelines execute several steps of data preprocessing, feature extraction and selection, and model execution in an iterative manner to improve the predictive performance of the machine learning algorithms.

4. RESULTS

To answer the research questions presented in Section 1, four separate studies were conducted. In the first study, all six data sets were combined and used to train and test the seven different machine algorithms that are explained in Section 3.4.4. The purpose of this study was (i) to observe if the machine learning algorithms could accurately predict the veracity of tweets from multiple different events, and thus allow analysts to monitor multiple cases of misinformation simultaneously, and (ii) to identify which algorithm performs best in predicting the veracity of tweets. This algorithm was then used throughout the remaining studies. In the second study, we test different partitions of training and testing data in order to analyze how the algorithm behaves with limited training data. In the third study, the six cases are investigated individually to analyze how the algorithm performs on single cases of misinformation. In the fourth and final study, five of the misinformation cases were used to train the model, and the sixth case was used to test. Data from the sixth case were

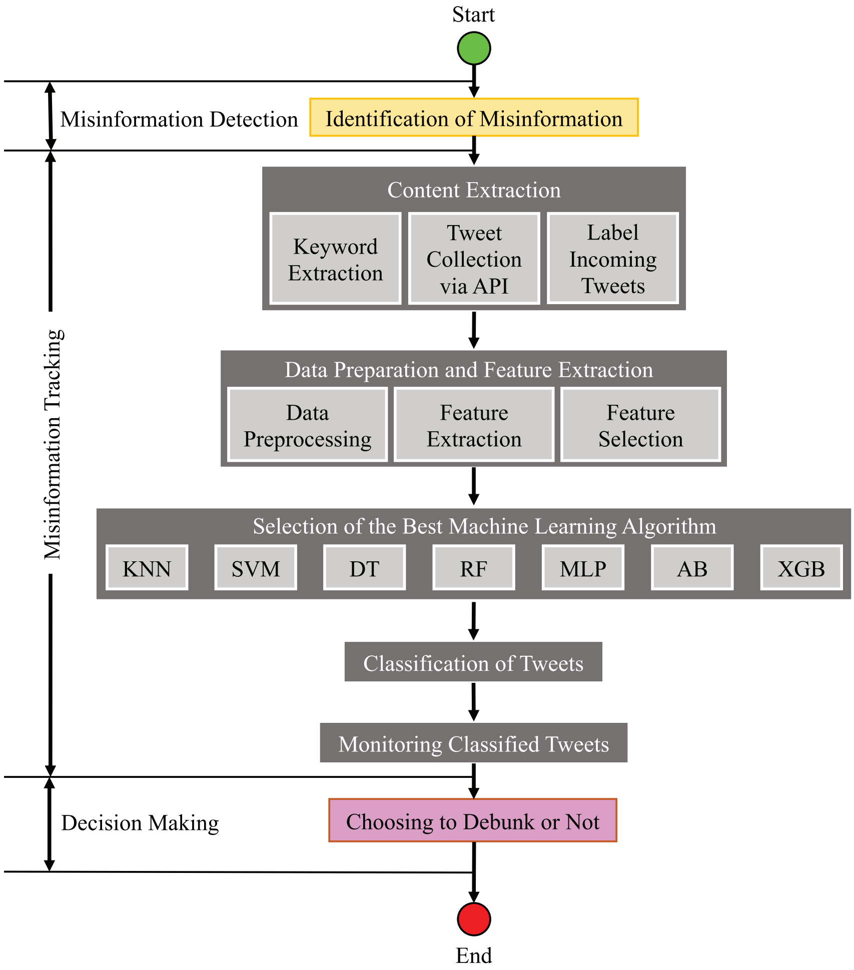

Fig 1. A flowchart showing the framework that is developed, implemented, and analyzed throughout this study.

slowly added to the training set until 90% of the sixth case was included in training. This study allowed us to analyze (i) if historical (labeled) data can be used to predict the veracity of new tweets, and (ii) if training the model on the five cases plus a portion of data from the sixth case increases the performance when predicting the veracity of the tweets in the sixth case. The results from all four of these studies are provided below. For each study, model training is always completed using fivefold stratified cross-validation to obtain the optimal combination of model hyperparameters.

4.1. Monitoring Multiple Events and Selecting the Best Performing Algorithm

In this study, all six of the data sets were combined, resulting in one master data set consisting of the total 15,952 tweets. After the tweets were randomly sorted, feature extraction and feature selection were carried out on the master data set. We then trained the algorithms on 80% of the data, and used the remaining 20% in order the test the performance of the seven algorithms. Given that multiple cases of misinformation often spread simultaneously on Twitter (FEMA, 2017a, 2017b), it is critical to understand if machine learning models can learn the feature space of multiple cases concurrently, and therefore accurately predict the veracity of tweets across multiple cases. The results for all seven algorithms are provided in Table VII.

From these results, we identify that the best performance comes from SVM, which achieved a macro-average F_1 score of 87.2%. This algorithm took 78.1 minutes to train, and proved to be very accurate in predicting the veracity of tweets across all. The worst performing algorithm was DT, which obtained a macro-average F_1 score of 78.8%, while taking just 9.7 minutes to train. The confusion matrices for both SVM and DT are provided in Appendix A, Tables A1

Table VII. Performance of the Different Algorithms in Predicting the Veracity of Tweets

ML Algorithms	Macro-Average Precision	Macro-Average Recall	Macro-Average **F**$_1$ Score	Training Runtime
kNN	82.3%	83.3%	82.7%	5.3 minutes
DT	78.4%	79.6%	78.8%	9.7 minutes
RF	86.0%	86.1%	86.0%	22.5 minutes
XGB	85.9%	86.3%	86.1%	202.0 minutes
AB	84.5%	84.4%	84.4%	10.0 minutes
SVM	87.1%	87.3%	87.2%	78.1 minutes
MLP	85.5%	85.7%	85.6%	140.7 minutes

Note: The best performance came from support vector machine (SVM), and the worst performance came from decision tree (DT).

and A2, respectively. These confusion matrices show that the framework is able to effectively learn the feature space of the entire data set in training, and therefore predict dominant and minority classes with high performance. Although DT obtained an F_1 score 78.8%, all of the other algorithms exceeded 82%, with four of them exceeding 85%. This implies that the algorithms can accurately and effectively process features from multiple different cases of misinformation, and perform very well in predicting the veracity of tweets. Given that multiple cases of misinformation often spread concurrently during disasters, these results prove to be a significant contribution to the literature and current practices.

Next, we analyzed the performance of the models in terms of model accuracy and count percentage as a function of cut-off probability. The output labels of each data point in the testing set are predicted in terms of probabilities. Let p_j be the cut-off probability, where $p_j \in [0, 1]$ and $j = 1, 2, \dots, n$. For p_j to assume discrete values, the following condition holds:

$$p_j = \{ \begin{matrix} p_{j-1} + \delta & \text{for} & j = 1, 2, \dots, n, \\ 0 & \text{for} & j = 0. \end{matrix} \qquad (7)$$

In Equation (7), δ is a predefined nonzero value between any two successive cut-off probabilities. Let c_j and a_j be the count and accuracy of predictions at cut-off probability p_j. We now define cumulative total count (S_j^c) and cumulative accuracy (S_j^a) based on c_j and a_j as follows:

$$S_j^c = \sum_{j'=j}^{n} c_j \quad \text{for} \quad j = 1, 2, \dots, n, \qquad (8)$$

$$S_j^a = \sum_{j'=j}^{n} a_j \quad \text{for} \quad j = 1, 2, \dots, n. \qquad (9)$$

Using Equations (8) and (9), we define model accuracy (r_j) and count percentage (q_j) at cut-off probability p_j as follows:

$$r_j = \frac{S_j^c}{S_j^a} \times 100\% \quad \text{for} \quad j = 1, 2, \dots, n, \qquad (10)$$

$$q_j = \frac{S_j^c}{\sum_{j=1}^{n} c_j} \times 100\% \quad \text{for} \quad j = 1, 2, \dots, n. \qquad (11)$$

In Fig. 2, we compare the performance of different models with respect to model accuracy and count percentage. With an increase in cut-off probability, there is a marginal improvement in model accuracy for DT and AB. This shows that DT and AB are unable to accurately predict unseen test cases even at high values of cut-off probabilities unlike models kNN, RF, XGB, SVM, and MLP that show a significant increase in model accuracy with increase in cut-off probability. For DT and AB, the change in count percentage with respect to cut-off probability is not as significant as compared to other models. SVM proves to be the best model in terms of the overall performance. With respect to model accuracy, its profile is similar to that of highly accurate models XGB, RF, kNN, and MLP. With respect to count percentage, as the cut-off probability is increased for XGB, MLP, SVM, kNN, and RF, the number of tweets whose prediction probabilities do not meet the threshold defined by cut-off probability increases significantly (as denoted by a decrease in the count percentage).

From these results, it is evident that although increasing the cut-off probability will significantly increase the model accuracy for kNN, RF, XGB, SVM, and MLP, it will decrease the count percentage. Therefore, there is a tradeoff in regards to the accuracy of these models and the number of data points whose prediction probabilities will meet the cut-off threshold.

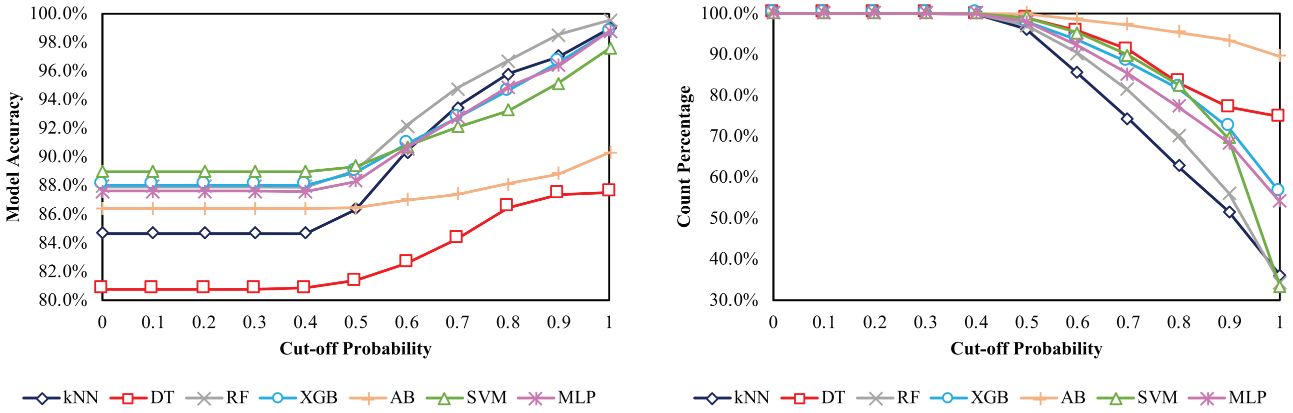

Fig 2. Model accuracy and count percentage as a function of cut-off probability.

Table VIII. Results from Different Data Partitions Using SVM

Train-Test Split (X–Y%)	Macro-Average Precision	Macro-Average Recall	Macro-Average F_1 Score	Training Runtime
90–10	86.9%	87.3%	87.1%	110.6 seconds
80–20	87.1%	87.3%	87.2%	94.9 seconds
70–30	86.1%	86.2%	86.2%	73.3 seconds
60–40	85.5%	85.8%	85.7%	52.6 seconds
50–50	86.0%	86.1%	86.0%	41.8 seconds
40–60	85.1%	84.9%	85.0%	29.0 seconds
30–70	84.2%	83.6%	83.9%	17.3 seconds
20–80	83.0%	82.3%	82.6%	10.8 seconds
10–90	79.6%	78.1%	78.7%	3.7 seconds

4.2. Data Partitioning

Upon identifying that the best performing algorithm was SVM, the next step was to analyze how the model performed with lower amounts of training data. In carrying out this study, the same extracted and selected features were used from Section 4.1. Given that annotating data can potentially take a lot of time, it is important to address the fact that the algorithm must continue to perform well with lower amounts of training data. Due to the time sensitivity when agencies track live cases of misinformation, the efforts put into labeling the data for model training must be minimized. This will allow the agencies to monitor and potentially debunk misinformation efficiently.

In this study, SVM was exposed to different amounts of training data. We first started by using 90% of the data in training (14,357 tweets) and decreased the size of the training data in 10% increments until finally reaching a training data set size of 10% of the total (1,595 tweets). The results are shown in Table VIII, and we note that SVM still performs well even when it is exposed to lower amounts of training data. In the case of using 10% of the data in training, SVM still achieves a 78.7% macro-average F_1 score, and trains in just 3.7 seconds. With just 20% of the data used in training, the model reaches an F_1 score of 82.6% in 10.8 seconds. The best performance comes from using 80% of the data to train, and 20% to test. In this case, SVM achieves an F_1 score of 87.2%, as reported in Section 4.1. Hence, we observed that when the algorithm was exposed to lower amounts of training data, its performance generally decreased along with the training time.

4.3. Monitoring Misinformation for Individual Events

After analyzing how SVM performs on all of the misinformation cases combined, the next task was to separate the cases and run the model on the individual events. For this study, feature extraction and

Table IX. Results from Predicting the Veracity of Tweets in the Individual Cases

Case No.	Event	Macro-Average Precision	Macro-Average Recall	Macro-Average F_1 Score	Training Runtime
1	Manchester Bombing Holiday Inn Rumor	66.8%	59.3%	62.2%	59.5 seconds
2	Hurricane Harvey Immigration Rumor	66.7%	56.9%	58.6%	47.5 seconds
3	Hurricane Irma Immigration Rumor	73.7%	73.6%	73.4%	11.8 seconds
4	Hawaii Missile False Alert	81.5%	77.3%	78.6%	389.8 seconds
5	Boston Marathon Bombing Sandy Hook Rumor	72.0%	64.9%	64.8%	116.0 seconds
6	Boston Marathon Bombing Donation Rumor	79.0%	76.1%	76.6%	16.4 seconds

Table X. F_1 Scores for Individual Classes of the Experimental Data Sets

		$\mathbf{F}_1$Score of Individual Class		
Case No.	Crisis Event	False	Neutral	True
1	Manchester Bombing Holiday Inn Rumor	97.6% (706)	12.5% (11)	76.4% (60)
2	Hurricane Harvey Immigration Rumor	23.1% (20)	63.3% (88)	89.5% (299)
3	Hurricane Irma Immigration Rumor	62.1% (15)	72.3% (38)	85.7% (66)
4	Hawaii Missile False Alert	59.4% (103)	82.3% (385)	94.1% (851)
5	Boston Marathon Bombing Sandy Hook Rumor	87.5% (193)	23.5% (53)	83.2% (174)
6	Boston Marathon Bombing Donation Rumor	87.1% (40)	60.0% (30)	82.6% (52)

Note: The number of testing data points corresponding to each class for a given event is enclosed within parentheses.

feature selection were carried out separately for every case (data set). For every data set, we used 80% of the data to train SVM, and subsequently tested the model with the remaining 20% of data. Given that there may be times when only one case of misinformation is propagating on Twitter, or authorities are only concerned with one case of misinformation, it is important to analyze how SVM performs on a single case. In Table IX, the results for all six cases are provided. Although the results for three of the cases (1, 2, and 5) indicated predictive performance below 65%, it can be observed that the other three cases (3, 4, and 6) obtained macro-average F_1 scores over 73%, reaching a maximum of 78.6% for the Hawaii missile false alert data set.

A deeper analysis on the predictive performance of SVM for individual classes is provided in Table X. As previously discussed in Section 3.4, the predictive performance of the classification system is evaluated using a macro-average F_1 score, which treats all of the classes equally. Therefore, a class with an extremely low F_1 score greatly reduces the overall macro-average F_1 score for the given data set. A direct observation of this is observed in cases 1, 2, and 5. As shown in Table X, in cases 1 and 5, the neutral class has a much lower F_1 score as compared to the other classes, while in case 2, it is the false class that suffers a lower F_1 score. The reason behind such low scores for these classes is primarily due to an extremely low number of data points in comparison to the other classes. This makes the learning process more difficult for algorithm, thereby resulting in attenuated overall performance for cases 1, 2, and 5.

A comparison between the predictive performance of SVM and a naïve classification system is provided in Table XI. The naïve classifier randomly predicts a class in the testing data set in proportion to the class distribution of the training data set. The classification strategy of the naïve classifier is entirely dependent on the base-rates for the three

Table XI. Comparison of Macro-Average F_1 Scores Using SVM and a Naïve Classifier

Case No.	Crisis Event	Macro-Average $\mathbf{F}_1$ Score using SVM Classifier (X)	Macro-Average $\mathbf{F}_1$ Score using Naïve Classifier (Y)	Difference in Macro-Average $\mathbf{F}_1$ Scores (X–Y)
1	Manchester Bombing Holiday Inn Rumor	62.2%	33.7%	28.5%
2	Hurricane Harvey Immigration Rumor	58.6%	35.1%	23.5%
3	Hurricane Irma Immigration Rumor	73.4%	31.0%	42.4%
4	Hawaii Missile False Alert	78.6%	31.8%	46.8%
5	Boston Marathon Bombing Sandy Hook Rumor	64.8%	35.0%	29.8%
6	Boston Marathon Bombing Donation Rumor	76.6%	33.7%	42.9%

classes (true, false, or neutral) in the training data set, without taking into consideration the content of the tweets. In Table XI, it is observed that the performance of SVM is far superior than the performance obtained by the naïve classifier across all the cases. This difference in performance between the two classifiers is much more significant (greater than 40%) in cases 3, 4, and 6 than in cases 1, 2, and 5, where the observed difference lies in the range of 20–30%. Therefore, we can conclude that with respect to the naïve classification system, SVM is quite powerful in predicting the veracity of tweets across individual cases.

4.4. Using a Corpus of Labeled Tweets to Track a New Case of Misinformation

From the results obtained in Section 4.3, it is clear that SVM is able to predict the veracity of tweets for single cases, although it performed better when being exposed to multiple cases in Sections 4.1 and 4.2. In this study, we consider the leave one out (LOO) setting (Lukasik et al., 2016), where the algorithm is trained on $n-1$ cases and its prediction performance is tested on the nth case. Results from this study allow us to understand if SVM can predict the veracity of tweets in a new, unseen case of misinformation by using a model that is trained on previous cases. Following random selection, we remove the Boston Marathon bombing donation rumor (referred to as the "donation rumor" throughout this section) from the combined data set, and use the remaining five cases to train SVM. In the first trial, we test the model on all of the data from the donation rumor to analyze how SVM performs in predicting a brand new case that it was unexposed to in training. Next, we slowly add the donation rumor data into the training set in 10% increments. The results from this allow us to analyze if SVM can obtain better predictive performance on the donation rumor case when training the model on the five cases plus a portion of data from the donation rumor case. The results from this study are provided in Table XII. From the results, it is observed that by using only the data from the five cases to predict the donation rumor case, SVM does not show high performance, signified by a macro-average F_1 score of 50.6%. When adding data from the donation rumor into the training set, the performance improves significantly, and with 40% of the donation rumor data added into the training set, we observe the maximum performance in this study, with an F_1 score of 73.4%.

Table XII. Results from Using Five Cases in the Training Phase, and the Sixth Case in the Testing Phase

Amount of Boston Bombing Donation Data in Training Set	Macro-Average Precision	Macro-Average Recall	Macro-Average $\mathbf{F}_1$ -Score
0%	71.2%	56.9%	50.6%
10%	68.6%	67.3%	67.1%
20%	68.7%	67.3%	66.9%
30%	73.6%	73.1%	72.4%
40%	74.8%	73.6%	73.4%
50%	70.7%	69.3%	69.0%
60%	70.8%	69.8%	69.8%
70%	67.3%	66.0%	66.2%
80%	69.8%	69.3%	69.4%
90%	73.1%	70.0%	69.0%

Note: Tweets from the sixth case were added to the training set in 10% intervals until finally 90% of the sixth case was used in training. SVM was the algorithm used.

In Table XIII, we provide the results from using the same data partitioning schema that was utilized in Section 4.2, except this time we only partition on

Table XIII. Results from Different Training and Testing Partitions on the Boston Marathon Bombing Donation Rumor Data

Train-Test Split(X–Y%)	Macro-Average Precision	Macro-Average Recall	Macro-Average **F**$_1$ -Score
10–90	81.4%	64.1%	58.4%
20–80	77.5%	68.4%	65.6%
30–70	75.4%	70.2%	68.7%
40–60	75.7%	72.3%	71.8%
50–50	75.6%	73.6%	73.4%
60–40	79.2%	76.8%	76.6%
70–30	84.0%	83.1%	83.3%
80–20	79.0%	76.1%	76.6%
90–10	77.7%	78.7%	77.7%

Note: SVM was the model used.

the donation rumor. By obtaining these results, we can compare Tables XII and XIII to understand if it is better to use data from other cases in training, or just the case of interest. Overall, it is clear from Table XIII that SVM was able to perform much better when using only the data from the donation rumor to train. When using 70% of the donation rumor data to train, and 30% to test, SVM achieved an F_1 score of 83.3%, which surpasses all of the F_1 scores presented in Table XII.

Comparing Tables XII and XIII, it can be observed that with 10–40% of the donation rumor data used in training, SVM performs best when also training on the data from the other five cases, but when ≥50% of the donation rumor data were used in training, SVM performs better when not adding the other five cases.

5. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

Given the danger and unneeded confusion that misinformation causes in current society, the framework and results that are presented throughout this research offer a timely and significant contribution to the related literature. Through the development and thorough analysis of a machine learning framework that can aid in the monitoring and control of misinformation, we offer agencies, decisionmakers, and researchers new knowledge and tools. Not only do these contributions support the resolution of false information during disasters, but they also offer researchers around the world with a methodology that can be used to automatically annotate large Twitter data sets or other related text-based data. Historically, the classification and annotation of Twitter data for research has been a manual task, requiring humans to code every single tweet in the data set. This task requires a significant number of hours and manpower (in our study, over 50 hours and three separate coders were needed). By deploying the framework that we provide, researchers can label a small percentage of their data, train a machine learning model such as SVM on that labeled data, and automatically predict the rest of the labels.

As a result of the four studies that were conducted in this research, we are able to draw many conclusions. In the first study, seven different machine learning algorithms were analyzed. Using all six of the data sets that were collected and coded, the seven algorithms were trained with 80% of the data (via fivefold stratified cross-validation), and tested on the remaining 20% of the data. The results from this study revealed that the best algorithm for the task of classifying the veracity of tweets was SVM, which achieved a macro-average F_1 score of 87.2%. The remaining six algorithms also exhibited high predictive performance, with the lowest F_1 score being 78.8% with DT, and all other algorithms exceeding 82%. From this study, we conclude that machine learning is a very good tool for automatically classifying the veracity of misinformation-related Twitter data. In addition, the algorithms were successful in predicting multiple different cases simultaneously. When increasing the cut-off probabilities, algorithms such as RF and SVM were able to obtain F_1 scores of almost 100%, signifying near perfect performance. The drawback of increasing the cut-off probability to ensure top performance is that some of the data points will not fall within any class, since their output probability for a specific class does not meet the cut-off probability.

From the results of the second study, it is evident that SVM continues to perform well without a significant amount of training data. With just 10% of the data used in training, SVM achieved an F_1 score of 78.7%. This is important for agencies and decisions makers, as their time spent labeling the training data needs to be minimized in time-sensitive situations. In the third study, SVM was trained on all of the six cases separately to analyze how the model performed in predicting the veracity of tweets in single cases. The results indicate a lower performance compared to using all six cases together. For the Hurricane Harvey immigration rumor, SVM achieved its lowest F_1

score of 58.6%, and for the Hawaii missile false alert, SVM achieved its highest F_1 score of 78.6%. These results suggest that SVM struggled to learn the features of some of the cases when trained individually, although in Table XI we were able to identify that SVM greatly outperforms a naïve classifier. These results and analyses motivated the idea for carrying out the fourth study.

In the fourth and final study, we aimed to predict the veracity of tweets in a "new" case by training SVM on previous cases. We were interested to see if the features and patterns identified from training the model on "historical" cases could be used to accurately predict a new case. The results from the first part of this study suggest that machine learning does not perform well in this context, indicated by a low F_1 score of 50.6%. This is likely due to the large variation in the textual features that were extracted between the cases. Since every case of misinformation is usually very unique in its content and topic, it is logical that it may be difficult to predict a new case based on the feature space from other cases. For the second part of this study, we added some of the data from the new case into the training set. The purpose of this was to understand if the combination of historical data and some new data in training would allow the model to learn better, and subsequently predict the new case with high performance. The results from this part of the study showed increased performance, with the model achieving a maximum F_1 score of 73.4% when including 40% of the new data in the training data set. These results indicate that the model was able to use the features from the historical cases, along with the small percentage of added features from the new case, in order perform well in predicting the veracity of tweets in the new case.

By successfully predicting the veracity of tweets that are spread during disasters, agencies and decisionmakers can utilize this framework to analyze how many Twitter users are spreading false information regarding a specific topic. Similarly, the concerned authorities can identify who the accounts are that are contributing to the spread of misinformation. If trusted or verified accounts are spreading misinformation, or if not many users have posted the truth, then the concerned authorities may choose to debunk the misinformation in order to clarify any confusion or rid any malicious intentions.

Future research in this domain should deploy and test this framework in a live scenario. For instance, following a disaster, researchers can collect the rumor-related tweets in real time, and label the veracity of a select amount of them. Using the framework provided in Fig. 1, machine learning algorithms can then be trained and deployed on unlabeled tweets in order to understand and analyze if there is significant misinformation being disseminated related to the topic.

In addition, future research should consider unsupervised techniques to accurately separate the data into different clusters. Research and developments in this domain could remove the need for labeling the data if unsupervised machine learning approaches could automatically create the different classes.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation (NSF) under Award No. 1762807. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We also thank the editors and two anonymous referees for providing constructive comments.

APPENDIX A: Confusion Matrices from the Best and Worst Performing Algorithms in Section 4.1

Table AI. The Confusion Matrix for the Best Performing Algorithm (SVM)

	Predicted			
Actual	False	Neutral	True	Total
False	952 (90%)	80 (8%)	26 (2%)	1,058 (100%)
Neutral	42 (7%)	511 (80%)	89 (14%)	642 (100%)
True	50 (3%)	65 (4%)	1,376 (92%)	1,491 (100%)
Total	1,044 (33%)	656 (21%)	1,491 (47%)	3,191 (100%)

Table AII. The Confusion Matrix for the Worst Performing Algorithm (DT)

	Predicted			
Actual	False	Neutral	True	Total
False	881 (83%)	105 (10%)	72 (7%)	1,058 (100%)
Neutral	54 (8%)	472 (74%)	116 (18%)	642 (100%)
True	86 (6%)	181 (12%)	1,224 (82%)	1,491 (100%)
Total	1,021 (32%)	758 (24%)	1,412 (44%)	3,191 (100%)

REFERENCES

Abedin, B., Babar, A., & Abbasi, A. (2014). Characterization of the use of social media in natural disasters: A systematic review. In *2014 IEEE Fourth International Conference on Big Data and Cloud Computing* (pp. 449–454).

Acar, A., & Muraki, Y. (2011). Twitter for crisis communication: Lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, *7*(3), 392–402.

Agarwal, P., Tang, J., Narayanan, A. N. L., & Zhuang, J. (2020). Big data and predictive analytics in fire risk using weather data. *Risk Analysis*, *40*(7), 1438–1449.

Arif, A., Shanahan, K., Chou, F.-J., Dosouto, Y., Starbird, K., & Spiro, E. S. (2016). How information snowballs: Exploring the role of exposure in online rumor propagation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 466–477).

Bergmeir, C., Costantini, M., & Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis*, *76*, 132–143.

Bruns, A., & Burgess, J. (2014). Crisis communication in natural disasters: The Queensland floods and Christchurch earthquakes. In *Twitter and Society*, Vol. 89 (pp. 373–384). New York: Peter Lang.

Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. *Internet Research*, *23*(5), 560–588.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, Third Annual Symposium on Document Analysis and Information Retrieval*, Vol. *161175*.

Cheng, J.-J., Liu, Y., Shen, B., & Yuan, W.-G. (2013). An epidemic model of rumor diffusion in online social networks. *European Physical Journal B*, *86*(1), Article No. 29. Retrieved from https://link.springer.com/article/10.1140%2Fepjb%2Fe2012-30483-5

Chua, A. Y. K., Tee, C.-Y., Pang, A., & Lim, E.-P. (2017). The retransmission of rumor and rumor correction messages on Twitter. *American Behavioral Scientist*, *61*(7), 707–723.

DHS. (2018). *Countering false information on social media in disasters and emergencies*. Technical report, U.S. Department of Homeland Security, Social Media Working Group for Emergency Services and Disaster Management. Retrieved from https://www.dhs.gov/sites/default/files/publications/SMWG_Countering-False-Info-Social-Media-Disasters-Emergencies_Mar2018-508.pdf

Dutta, H., Kwon, K. H., & Rao, H. R. (2018). A system for intergroup prejudice detection: The case of microblogging under terrorist attacks. *Decision Support Systems*, *113*, 11–21.

FEMA. (2017a). Harvey rumor control. Retrieved from https://www.fema.gov/disaster/4332/updates/rumor-control

FEMA. (2017b). Hurricane Irma rumor control. Retrieved January 2019 from https://www.fema.gov/hurricane-irma-rumor-control

Fischhoff, B., Gonzalez, R. M., Small, D. A., & Lerner, J. S. (2003). Evaluating the success of terror risk communications. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, *1*(4), 255–258.

Fraustino, J. D., Liu, B., & Jin, Y. (2012). *Social media use during disasters: A review of the knowledge base and gaps*. National Consortium for the Study of Terrorism and Responses to Terrorism [START].

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*, Vol. *1*. Springer Series in Statistics New York.

Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. In *Proceedings of the First Workshop on Privacy and Security in Online Social Media* (p. 2).

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hamidian, S., & Diab, M. T. (2015). Rumor detection and classification for Twitter data. In *Proceedings of the Fifth International Conference on Social Media Technologies, Communication, and Informatics* (pp. 71–77).

Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R., … Griffith, S. (2015). Social media and disasters: A functional framework for social media use in disaster planning, response, and research. *Disasters*, *39*(1), 1–22.

Hunt, K., Agarwal, P., Al Aziz, R., & Zhuang, J. (2020a). Fighting fake news during disasters. *OR/MS Today*, *47*(1), 34–39.

Hunt, K., Agarwal, P., & Zhuang, J. (2019a). A multi-algorithm approach for classifying misinformed Twitter data during crisis events. In *Proceedings of the 2019 IISE Annual Conference*.

Hunt, K., Agarwal, P., & Zhuang, J. (2019b). Tracking storms of misinformation spread amid disasters. *ISE Magazine*, *51*(9), 28–32.

Hunt, K., Wang, B., & Zhuang, J. (2020b). Misinformation debunking and cross-platform information sharing through Twitter during Hurricanes Harvey and Irma: A case study on shelters and ID checks. *Natural Hazards*, *101*(1), 861–883.

Jain, S., Sharma, V., & Kaushal, R. (2016). Towards automated real-time detection of misinformation on Twitter. In *2016 International Conference on Advances in Computing, Communications and Informatics* (pp. 2015–2020).

Jung, Y. (2018). Multiple predicting k-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, *30*(1), 197–215.

Kochkina, E., Liakata, M., & Augenstein, I. (2017). Turing at SemEval-2017 task 8, Sequential approach to rumour stance classification with branch-LSTM. *arXiv preprint arXiv:1704.07221*.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology*. Los Angeles, CA: Sage.

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining* (pp. 1103–1108).

Lazo, J. K., Bostrom, A., Morss, R. E., Demuth, J. L., & Lazrus, H. (2015). Factors affecting hurricane evacuation intentions. *Risk analysis*, *35*(10), 1837–1857.

Li, H., & Sakamoto, Y. (2015). Computing the veracity of information through crowds: A method for reducing the spread of false messages on social media. In *2015 48th Hawaii International Conference on System Sciences* (pp. 2003–2012).

Liao, Q., & Shi, L. (2013). She gets a sports car from our donation: Rumor transmission in a chinese microblogging community. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 587–598).

Lin, Y.-R., Margolin, D., & Wen, X. (2017). Tracking and analyzing individual distress following terrorist attacks using social media streams. *Risk Analysis*, *37*(8), 1580–1605.

Liu, H., & Motoda, H. (2012). *Feature selection for knowledge discovery and data mining*, Vol. 454. New York: Springer Science & Business Media.

Lukasik, M., Srijith, P., Vu, D., Bontcheva, K., Zubiaga, A., & Cohn, T. (2016). Hawkes processes for continuous time sequence classification: An application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 393–398).

Lundgren, R. E., & McMakin, A. H. (2018). *Risk communication: A handbook for communicating environmental, safety, and health risks*. Hoboken, NJ: Wiley.

Maddock, J., Starbird, K., Al-Hassani, H. J., Sandoval, D. E., Orand, M., & Mason, R. M. (2015a). Characterizing online rumoring behavior using multi-dimensional signatures. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 228–241).

Maddock, J., Starbird, K., & Mason, R. M. (2015b). Using historical Twitter data for research: Ethical challenges of tweet deletions. In *CSCW 2015 Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World*.

Maresh-Fuehrer, M. M., & Smith, R. (2016). Social media mapping innovations for crisis prevention, response, and evaluation. *Computers in Human Behavior*, *54*, 620–629.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: Can we trust what we RT? In *Proceedings of the First Workshop on Social Media Analytics* (pp. 71–79).

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

Murayama, Y., Saito, Y., & Nishioka, D. (2013). Trust issues in disaster communications. In *2013 46th Hawaii International Conference on System Sciences* (pp. 335–342).

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly*, *37*(2), 407–426.

Park, M., Allen, K., Davidson, L., & Turrell, L. (2018). Hawaii false missile alert "button pusher" is fired. Retrieved from https://www.cnn.com/2018/01/30/us/hawaii-false-alarm-investigation/index.html

Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: Methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, *16*(3), 197–214.

Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1589–1599).

Qiu, L. (2017). Fact check: Manchester bombing rumors and hoaxes. Retrieved https://www.nytimes.com/2017/05/24/world/europe/fact-check-manchester-bombing-rumors-and-hoaxes.html.

Sager, J. (2013). 10 Boston marathon bombing rumors that need to be stopped immediately. Retrieved from https://thestir.cafemom.com/crime/154242/10_boston_marathon_bombing_rumors

Shin, J., Jian, L., Driscoll, K., & Bar, F. (2017). Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *New Media & Society*, *19*(8), 1214–1235.

Simon, T., Goldberg, A., & Adini, B. (2015). Socializing in emergencies—A review of the use of social media in emergency situations. *International Journal of Information Management*, *35*(5), 609–619.

Song, C., Guo, C., Hunt, K., & Zhuang, J. (2020). An analysis of public opinions regarding take-away food safety: A 2015–2018 case study on Sina Weibo. *Foods*, *9*(4), 511.

Spence, P. R., Lachlan, K. A., Lin, X., & del Greco, M. (2015). Variability in Twitter content across the stages of a natural disaster: Implications for crisis communication. *Communication Quarterly*, *63*(2), 171–186.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. In *IConference 2014 Proceedings*.

Takayasu, M., Sato, K., Sano, Y., Yamada, K., Miura, W., & Takayasu, H. (2015). Rumor diffusion and convergence during the 3.11 earthquake: A Twitter case study. *PLoS One*, *10*(4), e0121443.

Tanaka, Y., Sakamoto, Y., & Matsuka, T. (2012). Transmission of rumor and criticism in Twitter after the great Japan earthquake. In *Annual Meeting of the Cognitive Science Society* (p. 2387).

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1079–1088).

Vos, S. C., Sutton, J., Yu, Y., Renshaw, S. L., Olson, M. K., Gibson, C. B., & Butts, C. T. (2018). Retweeting risk communication: The role of threat and efficacy. *Risk Analysis*, *38*(12), 2580–2598.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wang, B., & Zhuang, J. (2017). Crisis information distribution on Twitter: A content analysis of tweets during Hurricane Sandy. *Natural Hazards*, *89*(1), 161–181.

Wang, B., & Zhuang, J. (2018). Rumor response, debunking response, and decision makings of misinformed Twitter users during disasters. *Natural Hazards*, *93*(3), 1145–1162.

Xie, X.-F., Wang, M., Zhang, R.-G., Li, J., & Yu, Q.-Y. (2011). The role of emotions in risk communication. *Risk Analysis*, *31*(3), 450–465.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of ICML* (vol. 97, p. 35).

Yates, D., & Paquette, S. (2010). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem-Volume 47* (p. 42). American Society for Information Science.

Yuan, W., Guan, D., Huh, E.-N., & Lee, S. (2013). Harness human sensor networks for situational awareness in disaster reliefs: A survey. *IETE Technical Review*, *30*(3), 240–247.

Zeng, L., Starbird, K., & Spiro, E. S. (2016). # unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Tenth International AAAI Conference on Web and Social Media*.

Zhang, D., Wang, J., & Zhao, X. (2015a). Estimating the uncertainty of average F1 scores. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval* (pp. 317–320).

Zhang, Z., Zhang, Z., & Li, H. (2015b). Predictors of the authenticity of internet health rumours. *Health Information & Libraries Journal*, *32*(3), 195–205.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, *51*(2), Article No. 32. Retrieved from https://dl.acm.org/doi/10.1145/3161603

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., & Lukasik, M. (2016a). Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv:1609.09028*.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., & Tolmie, P. (2016b). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS One*, *11*(3), e0150989.