#Outage: Detecting Power and Communication Outages from Social Networks

Udit Paul
UC Santa Barbara
Santa Barbara, CA
u_paul@ucsb.edu

Alex Ermakov
UC Santa Barbara
Santa Barbara, CA
aermakov@ucsb.edu

Michael Nekrasov
UC Santa Barbara
Santa Barbara, CA
mnekrasov@cs.ucsb.edu

Vivek Adarsh
UC Santa Barbara
Santa Barbara, CA
vivek@cs.ucsb.edu

Elizabeth Belding
UC Santa Barbara
Santa Barbara, CA
ebelding@cs.ucsb.edu

ABSTRACT

Natural disasters are increasing worldwide at an alarming rate. To aid relief operations during and post disaster, humanitarian organizations rely on various types of situational information such as missing, trapped or injured people and damaged infrastructure in an area. Crucial and timely identification of infrastructure and utility damage is critical to properly plan and execute search and rescue operations. However, in the wake of natural disasters, real-time identification of this information becomes challenging. In this research, we investigate the use of tweets posted on the Twitter social media platform to detect power and communication outages during natural disasters. We first curate a data set of 18, 097 tweets based on domain-specific keywords obtained using Latent Dirichlet Allocation. We annotate the gathered data set to separate the tweets into different types of outage-related events: power outage, communication outage and both power-communication outage. We analyze the tweets to identify information such as popular words, length of words and hashtags as well as sentiments that are associated with tweets in these outage-related categories. Furthermore, we apply machine learning algorithms to classify these tweets into their respective categories. Our results show that simple classifiers such as the boosting algorithm are able to classify outage related tweets from unrelated tweets with close to 100% f1-score. Additionally, we observe that the transfer learning model, BERT, is able to classify different categories of outage-related tweets with close to 90% accuracy in less than 90 seconds of training and testing time, demonstrating that tweets can be mined in real-time to assist first responders during natural disasters.

CCS CONCEPTS

• **Information systems** → **Social networks**; **Information extraction**; **Clustering and classification**;

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380251

KEYWORDS

Event Detection, Information Extraction, Crisis Informatics, Natural Language Processing, Social Networks, Classification.

ACM Reference format:
Udit Paul, Alex Ermakov, Michael Nekrasov, Vivek Adarsh, and Elizabeth Belding. 2020. #Outage: Detecting Power and Communication Outages from Social Networks. In *Proceedings of Proceedings of The Web Conference 2020, Taipei, Taiwan, April 20–24, 2020 (WWW '20),* 11 pages.
https://doi.org/10.1145/3366423.3380251

1 INTRODUCTION

Users post content on social media platforms such as Twitter, Reddit and Facebook for a variety of purposes, including to report real-time situational incidents such as loss of electricity, internet connectivity and telecommunications [1]. During the onset of a natural disaster, situational information is posted by the affected individuals in real-time, including, increasingly, cries for assistance when 911 lines are overloaded [2]. First responders are responsible for carrying out rescue operations to help affected people during such emergency situations. Real-time social media posts can therefore provide critical information about the situation on the ground so that first responders can be most effective. Researchers have previously analyzed the usefulness of online information in timely crisis response and management [3, 4]. A key challenge is to extract valuable and actionable information such as missing or injured people and damaged utilities and infrastructure from all other content that appears online. It is therefore critical to develop information extraction tools that are capable of cutting through the noise and quickly filtering out vital information that authorities can use in their search and rescue operations.

Twitter has emerged as an ideal platform for information retrieval tasks due to the concise nature of the posts (tweets) [5]. Crisis informatics researchers have studied how to identify different types of sub-events, such as loss of lives and damage to infrastructure, from user generated posts [6, 7]. However, most of the developed algorithms focus on extracting information related to a wide spectrum of events, rather than a specific type of event [8, 9]. Since every type of event is not equally tweeted about by the users, some categories are classified with poor precision and recall as they represent only a small percentage of the entire dataset [10]. Additionally, in a recent study [11], it was reported

that the majority of the existing frameworks that aim to provide situational awareness to responders during a crisis do not meet the immediate informational requirements of specific responders. For example, information related to power outage would be more useful to responders responsible for restoring damaged utilities than responders in charge of locating trapped people. As such, there is an urgent need to develop highly domain specific information extraction tools to properly assist responders during emergencies.

In this work, we study the viability of the use of tweets to detect power and communication outages during natural disasters, with a specific focus on hurricanes. We begin by collecting tweets based on carefully selected keywords, and subsequently curate a *raw* dataset. We label a sample from the raw dataset to generate an *annotated* dataset that contains tweets related to power-outage, communication-outage and power-communication outage. Our goal is to first analyse characteristics, such as commonly used words, hashtags and sentiment, associated with the tweets that convey outage-related information during natural disasters. Then, we evaluate the performance of simple machine learning algorithms, neural network and transfer learning models to create a classification framework that is capable of determining whether or not a tweet is about outages. Once identified as an outage-related tweet, we perform an information extraction task to filter further information such as whether the tweet is about a power outage, a communication outage or both. To the best of our knowledge, no previous study has analyzed Twitter data in-depth to perform information extraction to detect both power and communication outages.

While prior work has shown that people often use Twitter as a platform to report power and communication outages [12], our study observes that over 75% of the tweets that contain outage-related keywords do not mention an actual outage. Hence, it is not enough to simply filter tweets based on keywords as this results in noisy dataset . Among tweets that actually mention an outage, we determine that the majority of tweets are made about power outages, followed by communication outages, and then both power and communication outages. Our analysis reveals that actual outage-related tweets carry more negative sentiment than tweets that contain outage-related keywords but that do not actually report an outage. As we attempt to classify these tweets, we are faced with the challenge of low numbers of usable tweets, as well as the inherent noise that is present in data gathered from Twitter. In spite of that, we observe that simpler models such as boosting and support vector machine are able to identify tweets that contain outage-related words with close to 100% accuracy. Furthermore, by applying state-of-the-art text classification techniques such as transfer learning, we are able to identify tweets that not only contain these keywords, but specifically report power and communication outages with very high accuracy, precision and recall scores. In summary, this work presents the following contributions:

- We curate a dataset of 18,097 unique tweets containing outage-related keywords, posted during seven major hurricanes that made landfall in the USA between 2012 and 2018.
- We present an in-depth analysis to determine features such as commonly used words, hashtags and sentiments associated with the tweets that mention power and communication outages.
- We use machine learning algorithms to perform multiple levels of information extraction to detect tweets that contain information about power and communication outages.
- We show that using simpler models such as SVM, tweets that contain outage-related keywords can be quickly detected with very high accuracy. Furthermore, employing transfer learning models such as BERT, we show that different types of outage-related events can be identified with high precision and recall scores.

The rest of the paper is organized as follows. We begin by reviewing related work that has been conducted in the areas of information extraction during crisis and emergency scenarios in Section 2. In Section 3, we introduce the dataset that we gathered and annotated for the purpose of this work. Section 4 presents information such as popular words, bi-grams, tri-grams and sentiment associated with various types of outage-related tweets. In Section 5, we explain the classification framework followed to first detect tweets that contain outage-related keywords and then classify different outage-related events. In Section 6, we present the results obtained by employing various machine learning algorithms to perform our classification tasks. Finally we conclude the paper in Section 7.

2 RELATED WORK

Information extraction from textual data is a very popular application of natural language processing. Previous work has been conducted to detect power outages using tweets as a source of information. Researchers have also focused on using data from social media to detect other types of events during natural disasters. In this section, we present the related work in two categories.

2.1 Power Outage Detection from Tweets

There has been some work that focused on detecting power outages using posts available on Twitter. In [13], the authors gathered a dataset and applied several machine learning algorithms to detect power outages from tweets. Their analysis showed that a multi-layer perceptron model is capable of detecting tweets related to power outages with reasonable accuracy, precision and recall. The authors in [12] used active learning, standard learning and Kleinberg's burst to detect real-time power outages using tweets. Supervised topic modelling was employed in [14] to detect power outages from tweets. Nightlight satellite imagery and tweets were used in [15] to identify locations of power outages. Specific keywords were used in [16] to gather a datset and then use classification algorithms to detect whether a tweet is about a power outage.

The primary focus of these studies was to make the binary distinction of whether or not a tweet refers to a power outage. Further, while these studies detect power outages from tweets using machine learning algorithms, they each utilized datasets that contained equal numbers of outage-related and unrelated tweets. In contrast, in this work, we maintain the ratio of tweets in each category that we observe during the analysis of our raw datset. Critically, in our

work, we only consider a tweet to be relevant to an outage if it mentions an actual outage and not simply if it contains outage-related keywords. Finally, in addition to identifying power outages, we also carry out detailed analysis to identify tweets that mention communication outages. To the best of our knowledge, our work is the first to perform detection of tweets that identify actual power and communication outages as well as to discern tweets that identify outages from tweets that simply contain outage-related words.

2.2 Sub-event Detection from Tweets

Prior work has attempted to identify information from social media during crisis scenarios [17, 18]. In [19], the authors attempt to use tweets to identify users in need of resources during or post natural disaster and match them with others who claim to have the needed resources. In [10], deep neural networks are used to identify useful tweets during crisis situations and categorize useful tweets. Tweets related to damaged infrastructure and utilities formed 8% of their dataset . The authors of [20] applied matching and learning based methods to identify tweets that provide situational awareness during natural disasters. In [21], the authors used integer linear programming to identify several types of events from tweets made during some natural disasters. In addition to natural disasters, sub-event detection from tweets has been performed in other fields. A recent paper [22] used keyword volume to identify specific events that belong to a category such as protests. The authors in [23] used tweets to gather information during epidemics. In each of these studies, different machine learning frameworks were employed to extract/classify information from a large number of data points gathered from Twitter. However, none of this work employed multiple levels of classification to obtain fine-grained information about power and communication outages.

3 DATA AND ANNOTATION

To achieve our goal of identification and classification of outage-related tweets, we first curate a raw dataset using specific keywords. Once we obtain the raw dataset, we manually perform annotation to generate an annotated dataset for detailed analysis and classification. Figure 1 presents our overall framework for this study.

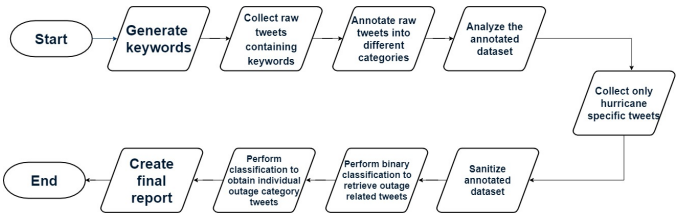


Figure 1: Proposed framework to detect power and/or communication outages from Tweets.

In this section, we describe our process to collect the raw dataset. We then explain the annotation procedure used to generate the annotated datset.

3.1 Dataset Curation

The volume of tweets related to infrastructure damage increases during natural disasters [24]. Unlike other natural disasters such as earthquakes that occur within a short span of time, hurricanes pass through an area over a much longer period, typically hours or days. As such, to curate our dataset, we selected the seven major hurricanes that made landfall in the USA between 2012 and 2018. We used Crimson Hexagon [25], a social media firehose with access to 100% of the Twitter stream, to collect tweets that appeared online in the time period from when each of these hurricanes made landfall to when they dissipated [26]. To collect tweets of interest, we generated two sets of keywords: Hurricane-specific keywords and Outage-specific keywords.

Hurricane-specific keywords: Similar to [21], to obtain hurricane-specific tweets, we filtered tweets using keywords such as, but not limited to, HurricaneMaria, harvey storm, hurricanematthew and #HurricaneSandy. In all cases, our keywords contained either the word "hurricane" or "storm", as well as the name of the storm. This resulted in a total of thirteen keywords per storm (91 total for the seven storms), each a permutation of these words and name combinations with different capitalization (i.e. we used each of michael storm, Michaelstorm, and MichaelStorm as a keyword). These formed our hurricane-specific keywords that we used to identify tweets related to these natural disasters.

Outage-specific keywords: In order to generate keywords to obtain tweets related to power and communication outages, we employed the semi-supervised topic modelling algorithm Latent Dirichlet Allocation (LDA) [27]. We began by scraping news articles that mention power and/or communication outages and formed a document containing the keywords mentioned in those articles. These keywords from the articles were obtained using the keywords class of the Newspaper3k [28] library provided by Python.To generate a diverse set of keywords, we applied LDA, with various combinations of numbers of topics and keywords, to this document. Five sets of topics, each having 15 keywords were heuristically determined to generate keywords of desired quality. Upon completion, we manually selected the words that we considered to be most relevant to obtain the required tweets. For example, keywords such as *blackout*, *outage*, *spotty*, *reception* and *damage* emerged from LDA as related keywords. We also added joined keywords such as *no power*, *can't call* and *call drop* to retrieve relevant tweets. Furthermore, to improve the quality of data, we collected tweets that had geo-location information and originated in the specific areas at the time the hurricanes passed through. The areas were determined from [26] and the geolocated tweets were collected using the location feature provided by Crimson Hexagon.

Table 1 presents the number of tweets that contained only hurricane-specific keywords as well as outage-specific keywords. The query containing outage-specific keywords also contained the hurricane-specific keywords for each hurricane. Among all tweets that have one or more hurricane-related keyword, only about $1-4$ percent of those also contain outage-related keywords. We observed that the overall number of tweets that contain tagged geolocation is on average 10 times less than the un-tagged tweets. Interestingly, the percentage of geo-tagged tweets that contain outage specific keywords in the total set is larger than those present in the un-tagged tweets.

Among the hurricanes, Hurricane Sandy contained the greatest number of tweets with outage-specific keywords by volume. This hurricane caused over 8 million people to lose power, far greater

Table 1: Number of tweets generated during hurricanes that contain keywords with and without geo-location.

Hurricane	Tweet extraction period	Keywords	Non-Geo-Tagged	Geo-tagged
Michael	10/06/2018-10/17/2018	hurricane-specific	387,617	62,191
		outage-specific	15,909	3,300
Florence	08/30/2018 -09/20/2018	hurricane-specific	718,414	69,262
		outage-specific	25,155	3,231
Maria	09/15/2017-10/03/2017	hurricane-specific	483,195	34,740
		outage-specific	26,509	1,594
Irma	08/29/2017-09/14/2017	hurricane-specific	1,761,869	252,082
		outage-specific	58,102	13,944
Harvey	08/16/2017-09/03/2017	hurricane-specific	1,372,863	193,965
		outage-specific	18,643	4,141
Matthew	09/28/2016-10/11/2016	hurricane-specific	1,202,774	175,941
		outage-specific	35,367	6,841
Sandy	10/22/2012-11/02/2012	hurricane-specific	1,903,552	250,936
		outage-specific	75,349	14,209

Table 2: Number of tweets per hurricane in the dataset.

Hurricane	Total Number of tweets selected
Michael	3,005
Florence	2,742
Maria	2,597
Irma	3,136
Harvey	1,208
Matthew	2,209
Sandy	3,200

than any other hurricane we studied [29]. Hurricane Maria also caused extensive power outages, leaving over 80,000 households without power [30]. In terms of communication outages, Hurricane Maria destroyed over 88% of the cell sites in Puerto Rico alone [31]. In comparison, cell phone infrastructure experienced less damage during Hurricane Sandy [32]. This could explain the larger number of outage-related tweets that appeared online during Hurricane Sandy than during Hurricane Maria; when faced with both cellular and power outages, many residents of Puerto Rico probably found themselves unable to post on Twitter. Hurricane Michael, on the other hand, had the fewest outage-related tweets, possibly because it also had the shortest duration among the hurricanes. In terms of percentage of outage-related tweets (percentage of tweets that contained outage keywords among all hurricane related tweets), Hurricane Maria contained the greatest number. When comparing geo-tagged tweets, we notice that Hurricane Sandy contained the greatest number of outage-related tweets, both by volume and percentage.

To curate our dataset, a sub-sample of the raw tweets that contained one or more of our outage-related keywords from each hurricane was selected. The sub-sampling strategy involved selecting a greater number of tweets that originated from the locations where the hurricanes made landfall. To ensure that our dataset was not dependent on one particular hurricane event, we incorporated roughly equal numbers of tweets from each hurricane. The smallest number of samples were drawn from Hurricane Harvey and Hurricane Matthew as they contained the smallest percentage of outage-related tweets (both geo-tagged and un-tagged) in their datasets. Table 2 presents the total number of tweets selected from

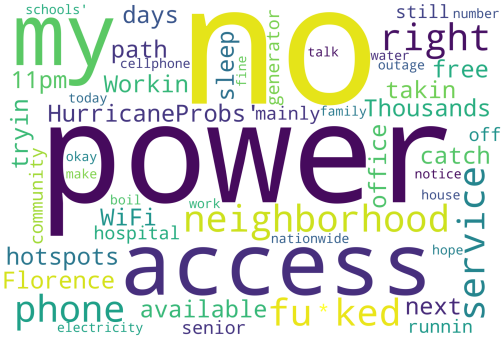


Figure 2: The salient words associated with power and communication outage tweets. A larger font for a word signifies high frequency of occurrence of that word in the dataset.

each hurricane. The gathered dataset consisted of 18,097 outage-related tweets. The salient words[1] present in the tweets is shown in Figure 2.

3.2 Dataset Annotation

To identify the different categories of tweets present in our raw dataset, we proceeded to annotate the dataset. We first attempted to perform the annotation process using Amazon Mechanical Turk (AMT) [33]. However, the annotated results obtained from AMT were unreliable; they contained many incorrectly labeled tweets, and in many cases multiple annotators labeled the same tweet differently. We therefore discarded these annotations. The annotation was then instead performed by 80 closely supervised volunteer upper division computer science students using Labelbox [34]. A pair of students were assigned the same subset of the *raw* dataset. The labels for the data points that did not match were further annotated by one of the authors. The annotators were provided detailed guidelines and asked to tag each tweet into one of the following four categories:

Not relevant: A large number of tweets in the raw dataset contained outage-related keywords but did not convey actionable outage-related information. For example, many tweets mention losing power in the future and thus do not provide any actionable information about current outages. As such, any tweets that do not contain current outage information are categorized as Not Relevant.

Power-outage: This category of tweets was reserved for tweets that mention power outages. In addition to directly reporting an outage, many tweets were informational in nature. They either contained a first-hand account by a person about a power outage in an area, or they contained a news article with information about areas currently experiencing an outage. Tweets that contained information about power restoration after a period of outage were also included in this category.

Communication-outage: Similar to the power-outage category, the category of communication-outage represents tweets that report communication outages. This category also consists of tweets that provided information about a related outage in an area/locality

[1] inappropriate language has been modified with the '*' character

Category	Example Tweet
Not Relevant	Hurricane Sandy - please don't take out my power and wifi.
Power-outage	Day 2 of no power..Thanks for everyone's concern.#HurricaneMatthew
Communication-outage	Internet just went out #hurricaneirma
Power-communication-outage	No access to my neighborhood right now no power and no phone service #f*cked #HurricaneSandy #HurricaneProbs

Figure 3: Example tweet per category.

Table 3: Number of annotated tweets per category.

Category	Number of Tweets
Not Relevant	13,957
Power-outage	2,791
Communication-outage	1,000
Power-communication-outage	349

as well as tweets that reported regaining communication facilities after an outage.

Power-Communication-outage: We observed a small number of tweets that mentioned both power and communication outages, and placed those tweets in this category. Note that tweets in this category do not necessarily indicate that both power and communication are out; instead, they provide information about the status of both utlity types.

Figure 3 shows an example tweet from each category; the total number of annotated tweets per category is presented in Table 3. Surprisingly, a large portion of the tweets belong to the Not Relevant class even though the tweets were carefully extracted using domain-specific keywords. The reason behind this is the tendency of people to use words such such *outage* and *blackout* to mention an anticipated outage in the future rather than using these words to report an active outage. Because we only annotated tweets about active outages in the outage-related categories, a large portion of the tweets ended up in the Not Relevant category.

4 DATASET ANALYSIS

In this section, we analyse the annotated dataset to better understand the nature of tweets that contain outage-related keywords. Our goal, through this analysis, is to highlight the differences that exist between the not-relevant class and others as well as between individual outage-related classes. In particular, we determine the inherent features such as popular words, bi-grams, tri-grams, hashtags and sentiments that are present in the tweets in each category. In Figure 4 we present the salient words associated with each of these four categories. We note that the not-relevant category consists of many of the same words that are present in other categories. However, as mentioned previously, the tweets in this category do not actually identify an outage. The salient words present in other categories are consistent with the names of the categories. Below we first perform lexical analysis to detect features such as single words, bi-grams, tri-grams and hashtags that are prevalent in each category. We then proceed to analyse the sentiments that are associated with the tweets by category.

4.1 Lexical Analysis

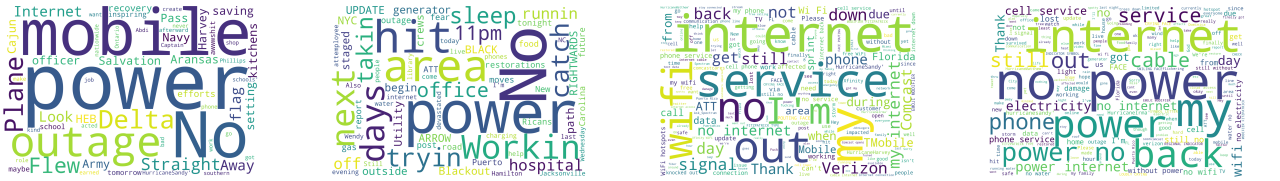
The lexical analysis of each category is presented below.

Not-relevant: This category consisted of over 75% of the total annotated tweets. In addition to investigating the most commonly occurring words in this category, as shown in Figure 4a, we evaluated the predominant bi-grams and tri-grams. Among single words, *power*, *mobile*, *No*, *outage* and *plane* were the five most frequent words in this category. For bi-grams, the words *my phone*, *no power*, *power outage*, *cell phone* and *phone call* appeared most frequently. From the frequently occurring words and bi-grams, it is not yet apparent that the tweets in this category do not convey any specific outage-related information. However, the most common tri-grams in this category, which include *get radio play*, *uncut internet station*, *charge my phone*, *got my phone* and *my phone off* shed more light on the nature of these tweets. Additionally, we investigate the frequent hashtags of the tweets from this category. The top three hashtags are *#mobile*, *#news* and *#tech*. Figure 5 shows the length of tweets in this category. On average, each tweet contained 18.6 words. We note the presence of a large number of outliers in the length of tweets in this category compared with others.

Power-outage: This is the second most popular category among the annotated data, containing 13% of the entire dataset. The category includes tweets that reported either power outages or restoration of power after an outage. The most frequently appearing words in this category are shown in Figure 4b. The top five common words are *No*, *power*, *hit*, *area* and *days*. Outage-related words such as *outage* and *blackout* also appear in the tweets in this category. The popular bi-grams in this group of tweets include *no power*, *still no*, *power outage*, *no electricity* and *without electricity*. Some of these words are also present in the commonly occurring tri-grams, which include *still no power*, *no power no*, *no power thanks*, *no power my* and *no power since*.

Further analysis of the 2,791 tweets in this category determined that 4% of the tweets mention power restoration after an outage. 20% of the power outage tweets were observed to be informative in nature, providing useful information about an outage. These informational tweets reported areas experiencing outages and in many cases included live updates from news organizations that stated the number of people experiencing outages in affected areas. The majority of tweets in this category, 76%, directly reported an outage during the time of the outage. Popular hashtags in this category are *#blackout*, *#nopower* and *#lightsout*. Figure 5 shows the distribution of the length of the tweets in this category. Tweets in this category contained 18.3 words on average.

Communication-outage: This category of tweets represents 5.5% of the overall number of tweets in our annotated dataset. Tweets in this category either inform about or report an active communication outage or mention having some form of communication capabilities returned after their loss. Popular words in this category include *internet*, *service*, *wifi*, *no* and *out* and are shown in Figure 4c. One interesting observation is that specific provider names, such as *verizon*, *tmobile* and *xfinity* appeared frequently in the tweets of



(a) Not-relevant **(b) Power-outage** **(c) Communication-outage** **(d) Power-communication-outage**

Figure 4: The salient words in each tweet category.

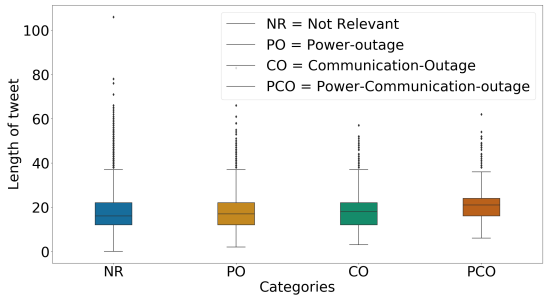


Figure 5: Length of tweets in each category.

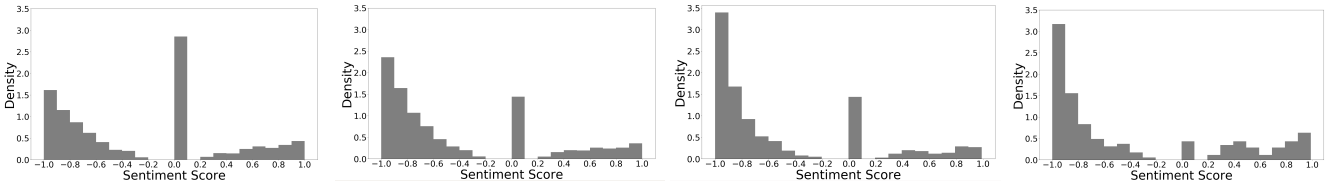
this category. This could be as a result of users being more familiar with the names of their telecommunication service providers. The most popular words pairs in the tweets of this category include *my internet*, *no internet*, *phone service*, *no service* and *internet down*. Tri-grams such as *cell phone service*, *my internet down*, *mobile networks knocked*, *networks knocked out* and *still no internet* emerged as the most common. The collection of these words indicate that when reporting a communication outage, people tend to use the word *service* together with *down*. Power outages are reported using *outage* and *blackout* in addition to *out*. Similar to the power-outage category, we subdivide the communication-outage related tweets into three subcategories. 9% of the tweets belong to the sub-category of tweets that mention restoration of communication service after an outage. 24% and 67% of the tweets in the communication-outage category inform or report about a communication outage, respectively. Hashtags *#wifi*, *#att* and *#internet* are the three most frequently used in this category. Unlike the most popular hashtags in the power-outage category, hashtags in this category do not inherently convey information related to an outage. Tweets in this category have an average length of 18.9 words as seen in Figure 5.

Power-communication-outage: This category contained the fewest tweets, about 2% of the overall annotated dataset. Because this category consists of tweets that must mention both power and communication outages, the average length of a tweet in this category, shown in Figure 5, is 22 words long. As can be seen from Figure 4d, these tweets combine the keywords from both the power-outage and communication-outage categories. Popular keywords include *power*, *internet*, *no*, *back* and *service*. Common bi-grams are *no power*, *power no*, *power internet*, *no internet* and *cell service*. We find that the tri-grams *still no power*, *power no cell*, *no cell service*, *no electricity no* and *no power internet* appear most frequently. As the tweets in this category are longer than the rest on average, we also determine the commonly occurring four-grams. These include *no power no cell*, *no power no internet*, *power no cell service*, *no power no wifi* and *no power cell service*. We observe that in addition to reporting about experiencing both power and communication outages, a number of tweets reported either having power but no communication or vice versa. Some tweets also provided information related to power and communication outages. When we analyze the nature of the tweets in this category further, we find that 10% of the tweets mention having power while experiencing some form of communication outage. Similarly, 10% of the tweets mention having communication capabilities while suffering from power outage. 15% of the tweets mentioned getting back both power and communication services after an outage. Informative tweets such as those providing locations and number of people experiencing power and communication outages formed 7% of the tweets in this category. Finally, 58% of the tweets reported experiencing both power and communication outages. The top three hashtags are *#poweroutage*, *#electricity* and *#finallygotpowerback*.

The lexical analysis of these categories highlights various salient features present in each category. The popular words and bi-grams of the not-relevant category are similar to those of the actual outage-related categories. In spite of the similarity between keywords, further analysis of the tri-grams and hashtags of these categories shows that the contents of the tweets in the not-relevant category do not report an outage. It is also noticed that during hurricanes, users tend to anticipate experiencing power and communication outages and post on Twitter before such outages actually occur. In addition to reporting an outage, tweets often mention restoration of services after an outage as well as provide meaningful information such as number of people experiencing an outage.

4.2 Sentiment Analysis

To better understand the inherent traits of the tweets that are present in these categories, we perform sentiment analysis [35] using the sentiment analysis API provided by IBM Watson [36]. IBM Watson analyzes the sentiment associated with a statement and assigns it a score between -1 and 1. A score closer to -1 conveys extremely negative sentiment while a score closer to 1 signifies more positive sentiment. Figure 6 shows the distribution of the sentiment scores of each of the four categories. The average sentiment score of not-relevant, power-outage, communication-outage and power-communication outage categories is calculated to be -0.26, -0.42, -0.51 and -0.40, respectively. As seen from Figure 6a, the sentiments associated with tweets in the not-relevant category are more neutral as they hover around 0. In contrast, the sentiment scores of the rest of the categories are more concentrated in the negative side of the scale with communication-outage tweets having the most negative sentiment. Overall, in the not-relevant



(a) Not-relevant (b) Power-outage (c) Communication-outage (d) Power-communication-outage

Figure 6: Distribution of sentiment scores of each category.

category, 29% of the tweets had sentiment score of 0, while 51% of the tweets attained negative scores. The percentage of tweets with negative sentiment score increased for the other categories. The power-outage, communication-outage and power-communication-outage categories had 68%, 72% and 70% of their tweets with score below 0, respectively.

5 OUTAGE-SPECIFIC CLASSIFICATION

In this section, we design a two-stage classification framework to automate the process of detecting outage-related tweets. Before performing the first level of classification, we collect a new set of tweets, using Crimson Hexagon, that occurred during the seven hurricanes. This dataset is comprised of tweets that contain only hurricane-specific keywords and not outage-specific keywords. These tweets are then added to the previously annotated dataset to form two separate classes of tweets. We first perform a binary classification to quickly extract all tweets that have our outage-specific keywords in addition to hurricane-specific keywords from the rest of the tweets. Before performing classification, we clean and pre-process the dataset. Once we have identified the outage-related tweets, we then perform the second level of classification, only on the annotated dataset, to automatically place tweets into the categories we established in the previous sections. Below we present the details associated with the pre-processing and classification tasks.

5.1 Pre-processing Dataset

Tweets are typically not properly grammatically structured and are likely to contain abbreviations, rendering them incomplete and noisy. In order to sanitize the dataset, we employed multiple text pre-processing steps. We removed URLs, non-ASCII characters and non-English characters. We also removed hashtags, user names and date and time strings. Emoticons were converted to UNICODE strings. To reduce the feature space, we converted all words to lower case.

We next created a custom set of stop words to ensure that we preserved the context of our tweets while eliminating unnecessary repeated stop words. For example, the stop words library provided by Python's NLTK contains 179 words such as *as*, *they*, *himself*, *out*, *down* and *not*. Removing the words *out*, *down*, *off*, *no* and *not* from our tweets could leave outage-related tweets meaningless. Hence we excluded these words from the stop words library. Additionally, we removed occurrences of event-specific words, such as hurricane, sandy and irma from the training dataset. This was done to ensure that the classifiers did not become dependent upon such words while identifying information that we require.

We used popular word embeddings frameworks to perform word vector initialisation. To generate word tokens, we first used term-frequency-inverse-document-frequency (tf-idf). Tf-idf is used to obtain the most important words within the tweets. These tokens from tf-idf were subsequently vectorized using GloVe [37]. We choose GloVe over another widely used word embedding framework, Word2Vec, due to the former's ability to take the ratio of the co-occurrence probabilities of consecutive words to establish semantic meanings for those words. For binary classification, we employed nine state-of-the-art classifiers such as logistic regression, support vector machine and K-nearest neighbors. These simpler classifiers were implemented using the scikit-learn 0.21 [38] library of Python. To extract various classes of our outage-related events, we used popular neural network models such as convolutional neural network (CNN) and recurrent neural network (RNN), in addition to the simpler models. These more sophisticated models were implemented using Keras with Tensorflow backend [39], as this platform contains the packages that are required to run these algorithms. Additionally, we implemented an emerging technique of text classification known as transfer learning to perform classification of our categories.

The classification task was carried out by splitting the overall training dataset into a ratio of 80 : 20 training to validation sets. All the classifiers were run on Google Cloud Compute powered by a 16GB NVIDIA Tesla V100 GPU. In addition to using a categorical cross entropy loss function with our neural network models, we also employed focal loss [40], which has been proven to be effective in classifying minority samples in image classification tasks. Next we present details associated with the two types of classifications we conducted and various methods we implemented to achieve better classification success.

5.2 Binary Classification

The goal of binary classification is to quickly isolate the domain-related tweets to conduct further information extraction. Specifically, we want to separate the tweets that contain hurricane-specific keywords from those that also contain outage-related keywords.

To perform binary classification, we first create a training dataset of a roughly equal number of samples that contain only hurricane-specific keywords (but not outage-specific keywords), comprising class 0, and tweets that contain both hurricane-specific and outage-specific keywords (our annotated dataset), forming class 1. We collected equal numbers of geo-tagged and un-tagged tweets that contained hurricane-specific keywords but excluded our outage-specific keywords using [25]. The training set consisted of 10,007 tweets, of which 5,236 samples belonged to class 0 and the rest to class 1. The distribution of the two classes in the test set, however,

was kept similar to what we observed while curating the original dataset in Section 3. Because outage-related tweets only comprised a very small fraction of the overall tweets that contained hurricane-related keywords, our test set contained 2,326 tweets, of which 2,203 belonged to class 0 and 123 belonged to class 1 (making up roughly 5% of the dataset). This small number of tweets of class 1 ensures consistency with what is observed during a real scenario. However, this results in difficulty in identifying these tweets with very high precision and recall. To perform this layer of classification, we only employed the simple classifier models as they are computationally inexpensive and capable of producing results with high accuracy.

5.3 Category Classification

Once we successfully filter tweets that contain outage-related keywords, we then attempt to further classify these tweets into the four major categories we established in Section 3. This is done to obtain more fine-grained information about different outage-related events. We first create a training set by selecting 3,500 random samples of the not-relevant class (class 0). The rest of the training set is formed of 2,295 randomly selected tweets from the power-outage category (class 1), 828 tweets from the communication-outage category (class 2) and 306 tweets from the power-communication outage category (class 3). As with the binary classification task, we kept the distribution of categories in the test set similar to the original dataset. In our test set, we selected 1,500 tweets from the not-relevant category, 496 tweets from the power-outage category, 172 from the communication-outage category and 43 tweets from the power-communication outage category.

In addition to the simple classifiers, we employed neural network and transfer learning models to extract tweets of each category in this layer of classification. To address the imbalance problem in our dataset, we applied the sampling technique SMOTE [41] and various sampling ratios amongst the classes. These techniques, however, fell short in improving the classification performance while detecting outage-related tweets, as they failed to adopt to the feature space that exists in our tweets. Therefore, because this is a multi-class classification problem, we instead first use a categorical-cross entropy loss function with a softmax layer in our neural network models. The categorical-cross entropy loss function can be mathematically defined as:

$$H(y, \hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N}(y_{ij})(log(\hat{y}_{ij}) \tag{1}$$

where H is the loss function, y is the actual label of the i^{th} observation of the j^{th} class and $\hat{y}$ is the predicted label for the observation made by the softmax layer of the neural network. An issue that arises with this loss function is that in a skewed dataset, it fails to properly penalise the classifier when it predicts the majority class. Because we are dealing with a dataset that exhibits class imbalance, we incorporate a loss function, known as focal loss, with our neural network classifier. Focal loss has proven to increase classification accuracy in datasets that suffer from the imbalance problem between classes [42]. Focal loss can be represented as:

$$FL(p_j) = \alpha(1 - p_j)^{\gamma} log(p_j) \tag{2}$$

where FL is the focal loss function and p_j is the softmax probability of the j^{th} class for a particular observation. α and γ are two regularizing parameters. This loss function adds more importance when the network predicts a minority sample as opposed to the overly represented sample. This makes it ideal for performing classification on an imbalanced dataset.

We choose a number of neural networks that have proven effective in text classification to perform this level of classification. To determine the ideal hyper-parameter configuration for each neural network, we use Grid Search [43] starting with multiple numbers of configurations. We train the CNN model using 100-word long embedding vectors alongside 512 convolutions filters of sizes 2, 3, 4, 5. To avoid over-fitting, we use a dropout of 0.5 while training with the Adam gradient descent optimizer [44]. The CNN model was run for 10 iterations with a batch size of 32. We also evaluated the performances of both LSTM and GRU-based bi-directional RNN. These RNN models were further incorporated with an attention layer to improve performance. We trained the RNN models containing 100 neurons for 20 iterations. We then employed Hierarchical Attention Network (HAN) [45] with 200 LSTM based word encoders and 250 sentence encoders. Finally, we tested the performance of Bidirectional Encoder Representations from Transformers (BERT) as a transfer learning model for the classification task [46]. Transfer learning models are pre-trained on a very large corpus and then fitted to perform classification on a smaller number of domain-specific data points. We used BERT-Large, Uncased (Original) model as the pre-trained model due to its ability to produce good results while remaining computationally inexpensive [47].

6 RESULTS

In this section we first present the results obtained after applying different classifiers to detect tweets that contain outage-related words. We then present the performance of the classification models in identifying specific outage categories. We compare the performance of the classification models by measuring the per-class precision, recall and f-score that each of these models produce. In addition, we compare the overall accuracy of each model as well as the time it takes for the model to perform the classification task. Because our goal is to classify the outage-related tweets quickly, the runtime for each algorithm presents us with important information we need to select the right model to perform the classification. Our goal is to determine the model that is able to quickly detect outage-related tweets with high accuracy, precision and recall scores.

6.1 Binary Classification

Table 4 presents the results we obtained after applying each of the nine classifiers on a curated dataset that contained only hurricane-specific keywords (class 0) as well as hurricane-specific and outage-related keywords (class 1). Almost every model performs exceptionally well in identifying tweets that contain only hurricane-specific keywords. The precision, recall and F1 scores of these models are very close to 1 when classifying members of class 0. In comparison, only a small set of models are able to identify samples of class 1 with good precision, recall and f-score. The boosting algorithm identifies class 1 tweets with the highest precision, recall and f-score values. Because the boosting algorithm has a hierarchical tree structure,

Table 4: Performance comparison of the binary classifiers.

	Class 0			Class 1				
Methods	Precision	Recall	F1-score	Precision	Recall	F1-score	Accuracy	Runtime(seconds)
Bagging	0.96	0.92	0.94	0.21	0.38	0.27	0.89	5.3
Boosting	0.99	1	1	0.99	0.94	0.97	0.99	1.94
Decision Trees	0.99	0.97	0.98	0.62	0.94	0.75	0.96	0.59
K-nearest neighbors	0.97	0.71	0.82	0.1	0.6	0.18	0.7	0.71
Logistic Regression	0.99	0.98	0.98	0.67	0.94	0.78	0.97	1.56
Multinomial Naive Bayes	0.99	0.87	0.93	0.28	0.9	0.43	0.87	0.12
Nearest Centroid	0.99	0.92	0.95	0.36	0.82	0.5	0.91	0.22
Random Forest	0.99	0.99	0.99	0.8	0.94	0.87	0.98	2.78
Support Vector Machine (SVM)	0.99	0.99	0.99	0.84	0.95	0.89	0.99	0.1

Table 5: Accuracy and runtime of the models used to perform outage-related categories classification.

Model	Accuracy	Runtime(Seconds)
Bagging	0.77	3.38
Boosting	0.84	9
Decision Trees	0.79	0.71
K-nearest neighbors	0.76	0.47
Logistic Regression	0.86	1.68
Multinomial Naive Bayes	0.82	0.11
Nearest Centroid	0.79	0.11
Random Forest	0.84	3.57
Support Vector Machine	0.86	0.16
CNN	0.65	645.05
CNN-Focal	0.84	650.89
RNN-LSTM	0.83	2309.47
RNN-GRU	0.84	1907.36
RNN-Attn-LSTM	0.84	2541.71
RNN-Attn-GRU	0.8	2216.57
RNN-LSTM-Focal	0.83	2239.26
RNN-GRU-Focal	0.83	1953.92
RNN-Attn-LSTM-Focal	0.83	2409.45
RNN-Attn-GRU-Focal	0.84	2261.31
HAN	0.85	2335.13
HAN-focal	0.82	2342.31
BERT	0.88	87

where a new tree learns from the results of the previously trained tree, it is able to perform better than other simple classifiers when performing binary classification. The SVM and random forest models achieve the second and third best performance in classifying the samples from class 1, respectively. K-nearest neighbor performs poorly when classifying class 1 samples. This occurred as a result of the insensitivity of the distance function of K-nearest neighbor towards small but meaningful differences between tweets. In addition to performing the overall classification task reasonably well, SVM also recorded the fastest run-time.

6.2 Outage-category Classification

Table 5 presents the accuracy and run-time of the classification models. Table 6 presents the classification performance achieved by these models in detecting not-relevant, power-outage, communication-outage and power-communication-outage tweets.

When comparing the accuracy of the models, Table 6 indicates that of the simpler models, boosting, logistic regression, random forest and SVM achieve accuracy scores above 0.8. These models also record low run-times, ranging from 0.16 to 9 seconds. The simpler models classify tweets from the not-relevant category with high precision and recall. In categorizing power-outage related tweets, the simpler models perform reasonably well, with logistic regression and boosting models achieving an f-score of 0.83. The performance of the simpler models, however, drops significantly while detecting samples from the two minority categories: communication-outage and power-communication-outage. This occurs as a result of these models' inability to learn classes with a small number of samples in an unbalanced dataset. Logistic regression, random forest and SVM are the only three models that produce an f-score greater than 0.50 when identifying tweets in the communication-outage category. In classifying tweets from the power-communication-outage category, among the simpler models, only the boosting and decision tree models achieve an f-score above 0.50.

As expected, the run-times of the neural network models are significantly greater than their simpler counterparts as shown in Table 5. CNN models execute fastest whereas RNN models take the longest. In terms of accuracy, except for the CNN model with categorical cross-entropy loss function, every other neural network model achieves accuracy scores around 0.80. The models also perform fairly similarly when classifying samples from the not-relevant categories. As seen in Table 6, the precision recorded by the neural network models exceeds 0.90 in detecting not-relevant tweets. Except for CNN with categorical cross-entropy loss function, all other models achieve recall scores of around 0.80 in detecting tweets of this class. Precision scores between 0.75 and 0.79 are reached by the neural network models when identifying power-outage tweets. The difference in performance between the simpler models and neural network models is seen when detecting communication and power-communication-outage tweets. The neural network models achieve higher recall scores in detecting communication-outage tweets while reaching better f-scores than the simpler models when identifying power-communication-outage tweets. Focal loss outperforms categorical-cross entropy loss when used with CNN across all four categories. It also records a 10% increase in f-score when used in conjunction with the RNN-LSTM model compared to the categorical cross-entropy loss function when detecting power-communication outage tweets.

Table 6: Classification performance of the models in detecting tweets per category.

Methods	Class 0			Class 1			Class 2			Class 3		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Bagging	0.84	0.87	0.86	0.65	0.72	0.68	0.47	0.31	0.37	0.33	0.09	0.14
Boosting	0.91	0.87	0.89	0.76	0.92	0.83	0.59	0.44	0.5	0.54	0.57	0.56
Decision Trees	0.9	0.8	0.85	0.73	0.87	0.79	0.41	0.56	0.47	0.45	0.61	0.52
K-nearest neighbors	0.85	0.85	0.85	0.65	0.69	0.67	0.41	0.39	0.4	0.21	0.09	0.13
Logistic Regression	0.92	0.9	0.91	0.77	0.92	0.83	0.63	0.52	0.57	0.65	0.25	0.36
Multinomial Naive Bayes	0.84	0.93	0.89	0.76	0.83	0.79	1	0.02	0.03	0	0	0
Nearest Centroid	0.92	0.8	0.86	0.8	0.82	0.81	0.34	0.62	0.44	0.34	0.68	0.45
Random Forest	0.91	0.88	0.9	0.72	0.93	0.82	0.66	0.47	0.55	0.69	0.2	0.32
Support Vector Machine	0.94	0.86	0.9	0.77	0.94	0.85	0.57	0.66	0.61	0.59	0.39	0.47
CNN	0.96	0.54	0.69	0.67	0.92	0.78	0.25	0.84	0.39	0.26	0.77	0.39
CNN-Focal	0.91	0.87	0.89	0.76	0.92	0.83	0.53	0.45	0.49	0.63	0.6	0.62
RNN-LSTM	0.95	0.81	0.87	0.76	0.91	0.83	0.49	0.79	0.6	0.48	0.7	0.57
RNN-GRU	0.93	0.84	0.88	0.75	0.91	0.82	0.56	0.64	0.6	0.63	0.79	0.7
RNN-Attn-LSTM	0.92	0.85	0.88	0.79	0.85	0.82	0.52	0.7	0.59	0.59	0.74	0.66
RNN-Attn-GRU	0.94	0.77	0.85	0.75	0.9	0.81	0.42	0.81	0.55	0.6	0.74	0.67
RNN-LSTM-Focal	0.93	0.82	0.87	0.76	0.89	0.82	0.48	0.73	0.58	0.56	0.84	0.67
RNN-GRU-Focal	0.94	0.81	0.87	0.76	0.92	0.83	0.48	0.74	0.59	0.59	0.79	0.67
RNN-Attn-LSTM-Focal	0.92	0.83	0.87	0.77	0.9	0.83	0.47	0.64	0.54	0.6	0.7	0.65
RNN-Attn-GRU-Focal	0.93	0.85	0.89	0.77	0.9	0.83	0.54	0.69	0.6	0.6	0.58	0.59
HAN	0.93	0.86	0.89	0.79	0.87	0.83	0.57	0.71	0.63	0.56	0.84	0.67
HAN-focal	0.96	0.77	0.86	0.74	0.96	0.84	0.46	0.77	0.57	0.47	0.88	0.61
BERT	0.93	0.9	0.91	0.83	0.89	0.86	0.67	0.66	0.66	0.69	0.84	0.7

The best performance in all the considered metrics is achieved by BERT in this classification task. From Table 5, we can see that though it takes longer to execute than the simpler models, it is able to achieve the highest accuracy; further, its run-time is faster than all the neural network models. It records the best f-scores when detecting tweets that belong to both the not-relevant class and outage-related categories. Because BERT is already pre-trained on a large corpus of texts, it is able to identify the tweets with very good performance, making it an ideal candidate to perform this classification task.

6.3 Remarks

With the aid of our annotated dataset and machine learning algorithms, we are able to detect outage-related events from a large stream of tweets that appeared online during recent hurricanes. The binary classifier is able to separate outage-specific tweets from others quickly, thereby reducing the amount of time needed to extract domain-specific tweets. Once these outage related tweets are detected, they can be further classified into different groups with the aid of an advanced learning model such as BERT. Using the ideal model to perform each level of information extraction will result in rapid classification of tweets, which first responders can then use to take immediate action or aid planning of additional operations.

7 CONCLUSION

In this work, we take an in-depth look at the tweets that originate during hurricanes and convey important outage-related information. We first determine keywords that are commonly used during power and/or communication outages. We use these keywords to obtain tweets that were posted during the seven major hurricanes in the USA between 2012 and 2018. These tweets were then annotated and placed into one of the four categories based on the outage information they contained. We perform a detailed analysis to better understand the type of tweets that belong to each of these categories. Finally, we apply various state-of-the-art machine learning models to first detect tweets that contain our outage-specific keywords and subsequently place them in their respective categories. Results show that computationally inexpensive models such as SVM and logistic regression can be used to filter out tweets that mention words related to outages. Through use of transfer learning models such as BERT, such outage-related tweets can be detected with high accuracy, precision and recall. Our framework can be implemented to provide first responders with outage related information during natural disasters. In our future work, we will build a user interface that incorporates classification models to perform real-time detection of outage-related tweets. Another next step is to conduct a deeper level of information extraction to sub-classify the tweets within each outage-related category. For example, with enough samples, we can train a classifier to automatically identify tweets that mention restoration of services or other useful information.

REFERENCES

[1] H. M. Saleem, F. A. Zamal, and D. Ruths. Tackling the challenges of situational awareness extraction in Twitter with an adaptive approach. *Procedia Engineering*, 107:301 – 311, 2015. Humanitarian Technology: Science, Systems and Global Impact 2015, HumTech2015.

[2] A. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6:248–260, February 2009.

[3] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.*, 47(4):1–38, June 2015.

[4] Y. Kryvasheyeu, H. Chen, N. Obradovich, E. Moro, P Van Hentenryck, J. Fowler, and M. Cebrian. Rapid assessment of disaster damage using social media activity. *Science Advances*, 2(3), March 2016.

[5] F. Alam, F. Ofli, M. Imran, and M. Aupetit. A Twitter tale of three hurricanes: Harvey, Irma, and Maria. In *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2018*, pages 553–572, May 2018.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931, April 2013.

[7] M. Imran, J. Castillo, C.and Lucas, P. Meier, and S. Vieweg. AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14. ACM, 2014.

[8] A. Schulz, E. Loza Mencía, T. Dang, and B. Schmidt. Evaluating multi-label classification of incident-related tweets. In *Proceedings, 4th Workshop on Making Sense of Microposts (#Microposts2014) at WWW: Big things come in small packages*, April 2014.

[9] K. Stowe, M. J. Paul, M. Palmer, L. Palen, and K. Anderson. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, SocialNLP@EMNLP 2016, Austin, TX, USA, November*, 2016.

[10] D. Nguyen, S. R. Joty, M. Imran, H. Sajjad, and P.t Mitra. Applications of online deep learning for crisis response using social media information. *CoRR*, abs/1610.01030, 2016.

[11] H. Zade, K. Shah, V. Rangarajan, P. Kshirsagar, M. Imran, and K. Starbird. From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *Proc. ACM Hum.-Comput. Interact.*, 2:1–18, November 2018.

[12] K. Bauman, A. Tuzhilin, and R. Zaczynski. Virtual power outage detection using social sensors. In *NYU Working Paper*, Sept 2015.

[13] S. S. Khan and J. Wei. Real-time power outage detection system using social sensing and neural networks. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 927–931, Nov 2018.

[14] H. Sun, Z. Wang, J. Wang, Z. Huang, N. Carrington, and J. Liao. Data-driven power outage detection by social sensors. *IEEE Transactions on Smart Grid*, 7(5):2516–2524, Sep. 2016.

[15] C. Hultquist, M. Simpson, G. Cervone, and Q. Huang. Using nightlight remote sensing imagery and twitter data to study power outages. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management*, pages 1–6, November 2015.

[16] K. Bauman, A. Tuzhilin, and R. Zaczynski. Using social sensors for detecting emergency events: A case of power outages in the electrical utility industry. *ACM Trans. Manage. Inf. Syst.*, 8:7:1–7:20, June 2017.

[17] L. Palen and A. L. Hughes. *Social Media in Disaster Communication*. Springer, 2018.

[18] C. Reuter, A. L Hughes, and M. Kaufhold. Social media in crisis management: An evaluation and analysis of crisis informatics research. *International Journal of Human–Computer Interaction*, 34(4):280–294, 2018.

[19] M. Basu, A. Shandilya, K. Ghosh, and S. Ghosh. Automatic matching of resource needs and availabilities in microblogs for post-disaster relief. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 25–26, Republic and Canton of Geneva, Switzerland, 2018.

[20] H. To, S. Agrawal, S. H. Kim, and C. Shahabi. On identifying disaster-related tweets: Matching-based or learning-based? In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, April 2017.

[21] K. Rudra, P. Goyal, N. Ganguly, P. Mitra, and M. Imran. Identifying sub-events and summarizing disaster-related information from microblogs. In *The 41st International ACM SIGIR Conference on Research; Development in Information Retrieval*, SIGIR '18, New York, NY, USA, 2018. ACM.

[22] A. H. Hossny and L. Mitchell. Event detection in Twitter: A keyword volume approach. *CoRR*, abs/1901.00570, 2019.

[23] K. Rudra, A. Sharma, N. Ganguly, and M. Imran. Classifying information from microblogs during epidemics. In *Proceedings of the 2017 International Conference on Digital Health*, DH '17, New York, NY, USA, 2017. ACM.

[24] M. Imran, P. Mitra, and C. Castillo. Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. *CoRR*, abs/1605.05894, 2016.

[25] Crimson Hexagon. https://www.brandwatch.com/#from-ch.

[26] List of United States hurricanes — Wikipedia, the free encyclopedia, 2019. [Online; accessed 17-September-2019].

[27] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, Jun 2019.

[28] Newspaper3k. https://newspaper.readthedocs.io/en/latest/.

[29] D. Henry and J. E. Ramirez-Marquez. On the impacts of power outages during hurricane sandy-a resilience-based analysis. *Systems Engineering*, 19(1):59–75, 2016.

[30] Hurricane Maria — Wikipedia, the free encyclopedia, 2019. [Online; accessed 23-September-2019].

[31] FCC. In *Communications Status Report for Areas Impacted by Hurricane Maria*. FCC, 2017.

[32] In Sandy's wake, here's why millions of Americans have cell service but no power, 2012. [Online; accessed 23-September-2019].

[33] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[34] Labelbox. https://labelbox.com.

[35] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 2008.

[36] Ibm watson. https://www.ibm.com/watson/services/natural-language-understanding.

[37] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014. Association for Computational Linguistics.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, November 2011.

[39] Keras. https://github.com/keras-team/keras.

[40] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[41] N. V. Chawla, Bowyer K. W., Hall L. O, and Kegelmeyer W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[42] R. Qin, K. Qiao, L. Wang, L. Zeng, J. Chen, and B. Yan. Weighted focal loss: An effective loss function to overcome unbalance problem of chest X-ray. *IOP Conference Series: Materials Science and Engineering*, 428:012022, Oct 2018.

[43] Tuning the hyper-parameters of an estimator. [Online; accessed 23-September-2019].

[44] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

[45] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[46] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[47] Tensorflow code and pre-trained models for BERT. https://github.com/google-research/bert.