Summer 2024

Harnessing Social Media for Disaster Response: Intelligent Identification of Reliable Rescue Requests During Hurricanes

Wael Khallouli
Old Dominion University, waelkhallouli@gmail.com

HARNESSING SOCIAL MEDIA FOR DISASTER RESPONSE: INTELLIGENT IDENTIFICATION OF RELIABLE RESCUE REQUESTS DURING HURRICANES

by

Wael Khallouli
M.S. in Computer Science, December 2013, University of Tunis, Tunisia
B.S. in Computer Science, December 2011, University of Tunis, Tunisia

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

ENGINEERING MANAGEMENT AND SYSTEMS ENGINEERING

OLD DOMINION UNIVERSITY
August 2024

Approved by:

Samuel Kovacic (Director)

Ghaith Rabadi (Member)

Andres Sousa-Poza (Member)

Jiang Li (Member)

ABSTRACT

HARNESSING SOCIAL MEDIA FOR DISASTER RESPONSE:
INTELLIGENT IDENTIFICATION OF RELIABLE RESCUE REQUESTS
DURING HURRICANES

Wael Khallouli
Old Dominion University, 2024
Director: Dr. Samuel Kovacic

Hurricanes pose a significant threat to both human lives and infrastructure. Decision-makers face substantial challenges during such events, as they must act quickly to address victims' needs. Social media platforms provide a valuable source for quick and real-time information. Recent hurricane events have shown that people turn to social media to call for help when official communication channels, such as 911, are overwhelmed. However, extracting actionable information from the massive number of messages posted on social media is challenging. Furthermore, verifying social media messages posted by the public is a critical concern for disaster response practitioners, making them hesitant to use this information. This study tackles the problem of identifying and assessing the reliability of actionable rescue messages posted on Twitter during hurricanes. A novel deep learning model is proposed for identifying rescue tweets, integrating a fine-tuned BERT model to extract low-level statistical features from the text and rule-based Regex filters to extract problem-specific features. In addition, a rule-based scoring model is introduced to assess the reliability of the identified rescue messages using a set of reliability indicators derived from the literature. The proposed models were evaluated using data collected and annotated from various hurricane events. The results indicated that the proposed classification model for identifying rescue tweets provides more robust results compared to

previous classification methods. Evaluated on rescue tweets from Hurricane Harvey, the proposed reliability assessment model could effectively identify reliable rescue tweets. The models developed in this study aim to improve the quality of actionable rescue information extracted from social media during hurricane events, enabling first responders to effectively integrate social media channels as a supplementary source of information in their decision-making process.

This dissertation is dedicated to my family and to
the loving memory of my late father.

ACKNOWLEDGMENTS

The completion of this dissertation would not have been possible without the support, guidance, and encouragement of many individuals. I would like to take this opportunity to express my deepest gratitude to them.

First and foremost, I am deeply grateful to my family. To my mom, my brother and his wife, my sister and her husband, my beloved fiancée, my dear nephews and nieces, my uncles, aunts, cousins, and all my relatives back in Tunisia: thank you for the unwavering support and encouragement you have given me over the years.

I would like to express my profound gratitude to my advisor, Dr. Samuel Kovacic, as well as my dissertation committee members, Dr. Ghaith Rabadi, Dr. Jiang Li, and Dr. Andres Sousa-Poza, for their valuable guidance, constructive discussions, and continuous belief in me. Thank you for everything.

I would also like to extend my gratitude to the faculty and staff of the EMSE department, especially Dr. Collins, Dr. Pazos, Dr. Handley, Dr. Cotter, Dr. Unal, Dr. Keating, and Dr. Bullington, for their support over the past few years. I would like to thank my former advisor, Dr. Huang, for his guidance during my early years in the EMSE department.

Lastly, I am deeply thankful to all my colleagues and dear friends in the EMSE and ECE departments, as well as my friends in Norfolk and across the U.S. Your support and encouragement have meant the world to me.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Natural disasters, such as hurricanes, earthquakes, and tsunamis, pose a significant threat to human lives and infrastructure. Over the years, their frequency and severity have increased [100]. Hurricanes are among the costliest natural disasters in the United States, causing billions of dollars in damages [128]. For instance, Hurricane Katrina (2005), one of the deadliest hurricanes in US history, caused significant damage to Mississippi and Louisiana. Eighty percent of the city of New Orleans was flooded, and 1,800 people lost their lives [17]. The total cost of Hurricane Katrina was estimated to be 172.5 billion (2021 USD) [128]. Similarly, Hurricane Harvey (2017) struck the southwest coast of Texas, leading to the loss of at least 88 lives and causing damages worth 133.8 billion (2021 USD) [133]. Hurricane Maria (2017) hit the northeastern Caribbean, resulting in the loss of tens of lives and damages worth 96.3 billion (2021 USD) [128]. As hurricanes become increasingly disastrous, improving current emergency response and mitigation strategies is crucial to alleviate their catastrophic impacts.

Disaster management is the research area that focuses on improving decision-making across the various stages of a disaster, including preparedness, response, recovery, and mitigation. Due to the rapid advancement of information and communication technology (ICT), disaster informatics [103] –a subfield of disaster management– has emerged as a field of study dedicated to studying how these technologies can improve disaster management processes. Among ICT technologies,

social media platforms have received considerable attention. A significant part of disaster informatics research has focused on exploring efficient methods to collect, process, and utilize social media data to improve decision-making during natural disasters.

Humanitarian relief organizations and first responders face numerous challenges during hurricanes due to time constraints, uncertainty, and limited availability of resources. Under conditions of extreme uncertainty, they must quickly and effectively allocate resources to mitigate the hurricanes' impacts, especially when human lives are at risk. Consequently, timely information is essential for an effective response to hurricane events. Solely relying on traditional and verified sources of information may not always be sufficient to meet first responders' needs [42]. Social media platforms provide a unique opportunity to gain access to large amounts of instantaneous data posted during hurricanes. However, extracting information from these platforms presents a significant challenge. Hurricanes trigger a sudden surge in communication, resulting in complex information scenarios and vast amounts of data, of which only a small portion is relevant to first responders [91]. Social media messages vary widely in usefulness, ranging from prayers and emotional support to incident reporting, calls for help, and situational updates. This diversity makes it difficult for disaster responders to filter useful information to inform their decision-making processes. This problem is known as the 'information overload' problem.

To address the 'information overload' problem, a wide range of research studies have developed computational methods for extracting useful information from the substantial volume of noisy data posted on social media platforms during natural disasters, including hurricanes. The primary goal was to enhance the 'quality' of information derived from social media and better

fulfill disaster responders' needs. A common approach involves training machine learning models to automatically categorize social media messages into different information types. Researchers have collected data from various disaster events and labeled them using several predefined categories, such as the relevance of the information to the disaster (e.g., relevant vs non-relevant, on-topic vs off-topic, among others) and humanitarian categories (e.g., causalities, caution and advice, infrastructure damage, among others), to train and evaluate the proposed classification systems. One significant weakness of the existing approaches is their focus on a general concept of 'situational awareness' that provide high-level information, which often fails to meet the precise needs of disaster responders for 'actionable' information. Consequently, Zade et al. [144] pointed out that, despite current advancements in developing social media information extraction methods, social media data streams are not well integrated into the formal disaster response workflow. To enhance the usability of social media channels, Zade et al. [144], Garcia et al. [42], Coche et al [29], and few other researchers have explored the concept of 'actionability' as a framework for building efficient systems to process social media data. Actionable information was defined by disaster response practitioners as any piece of information on social media they could use to assist, enact, and expedite action to an identified issue [144] (in other words, information that can trigger immediate action by first responders). Typical examples of actionable information are the implicit and explicit requests posted on social media during disasters, such as urgent rescue requests. The literature lacks research studies proposing effective methods for extracting actionable information. Rescue requests posted on social media platforms during hurricanes are a particular type of actionable information that needs further research.

1.2 RESEARCH PROBLEM

Urgent rescue messages posted on social media during natural disasters, such as hurricanes, are invaluable to search and rescue teams. These messages often contain details about individuals in need of immediate assistance, including their location, the nature of their urgent situation, and their needs. However, extracting these rescue messages is particularly challenging. Firstly, rescue messages are relatively rare and often hidden by a large volume of non-informative messages [137]. Secondly, during disasters, individuals tend to compose rescue messages in various informal ways, making it difficult to filter these messages by keywords. Furthermore, keyword-matching search approaches are inefficient and time-consuming. Therefore, developing efficient methods to automatically extract rescue-seeking messages from social media is of paramount importance. Given the critical nature of rescue information during natural disasters, it is crucial to continue improving current rescue identification methods for a better quality of rescue information.

For first responders, the reliability of information is a major concern [131]. Misinformation, rumors, and inaccurate information can rapidly spread through social media channels, which makes the process of evaluating the reliability of posted content more difficult [84]. Assessing whether information on social media is trustworthy enough to act upon during an emergency is challenging. Determining whether social media information is trustworthy enough to act upon during an emergency is challenging. Numerous studies, such as [144], [29], and [50], have highlighted the reliability issue of social media data through surveys and interviews with disaster response officials. A common challenge identified by these studies is the absence of adequate tools for verifying social media information. Reliability is a key component of actionable information on social media [144].

This dissertation investigates the problem of identifying actionable rescue-seeking messages posted on Twitter during hurricanes. Twitter, now rebranded as 'X,' was selected as a representative social media platform for the current study. For consistency, the name 'Twitter' will be used throughout this dissertation, as it was the platform's name at the time of data collection. Messages posted on Twitter will be referred to as 'tweets.' Natural disasters vary greatly in nature and intensity. The characteristics of rescue messages may differ depending on the type of disaster. Hence, this dissertation focuses on rescue tweets posted during hurricane disasters, leaving the analysis of other natural disaster types for future research. This problem is divided into two parts: the first part focuses on automatically identifying urgent rescue-seeking tweets (with location information), while the second part focuses on assessing their reliability.

Part 1: Identifying emergency rescue requests from Twitter during Hurricanes. Despite its practical relevance, research on identifying rescue requests posted on social media during natural disasters has received little attention [34]. This problem has been investigated in a few studies, such as [34] [148] and [137], where the authors trained machine learning methods to automatically identify rescue requests posted on Twitter during Hurricane Harvey. Additionally, [149] conducted a spatial and temporal analysis of rescue requests posted during Hurricane Harvey. However, previous studies have several limitations. Firstly, they have focused on a single disaster event (Hurricane Harvey) to conduct their analyses. Secondly, previous research studies have employed various machine learning and deep learning techniques, such as convolutional neural networks, support vector machines, and transformer-based models, to learn textual features from raw data for classification, but none of these models have investigated domain-specific features (i.e., specific textual patterns that match the characteristics of the rescue

messages posted on social media). It is unclear whether incorporating these features would enhance the predictive capacity of the proposed learning models, yielding more accurate identification of rescue messages. The first part addresses the following research questions:

- What are the main textual features that characterize emergency rescue request messages posted on Twitter during hurricanes?
- Using these features, how can existing models proposed for the rescue identification problem be improved?

Part 2: Assessing the reliability of rescue-seeking messages on Twitter during hurricanes.

The second part of this dissertation focuses on assessing the reliability of the identified rescue messages. While the credibility of online data, particularly social media data, has been extensively addressed in domains such as news, politics, and healthcare, it remains largely underexplored in the context of natural disasters and humanitarian disaster response. Only a few studies have addressed the social media data reliability problem in the disaster context. Examples include [53], [124], [139], and [16]. Although the reliability of social media information presents a key component of actionable information, to the best of the author's knowledge, none of these studies has addressed how to estimate the reliability of actionable social media information, including rescue messages posted on social media during hurricanes. Additionally, previous research studies proposing classification techniques for identifying emergency rescue tweets (e.g., [148], [34], [137]) have not examined the reliability of the identified rescue information. The second part addresses the following research questions:

- What are the key indicators for assessing the credibility of rescue-seeking messages posted on social media during natural disasters?
- How can the reliability of these rescue-seeking messages be estimated using these indicators?

1.3 RESEARCH GOAL AND CONTRIBUTIONS

The overarching goal of this dissertation is as follows:

To improve the usability of social media data streams in disaster response by enhancing existing models for extracting reliable and actionable rescue information.

In the first part, this dissertation introduces two novel classification models designed to identify rescue messages posted on Twitter during hurricanes **(first contribution)**. The first model employs a logic-based approach that uses regular expressions *regex* to catch the language patterns or features of rescue tweets. The second model is a novel deep learning-based framework that combines a state-of-the-art deep learning model for text classification (BERT) for low-level feature extraction and rule-based regex filters for high-level feature extraction within a single architecture.

In the second part, this dissertation introduces a two-stage rule-based reliability scoring system to evaluate the reliability of rescue messages posted on Twitter during hurricanes **(second contribution)**. This model integrates reliability indicators (factors) across multiple assessment dimensions, including user-level, content-level, and context-level assessment.

1.4 SIGNIFICANCE

The proposed rescue detection framework is illustrated in Figure 1. This research is motivated by observations from Hurricane Harvey. During Hurricane Harvey and its subsequent flooding, many stranded individuals were unable to reach 911 and other emergency centers due to the over-whelming surge in emergency calls that exceeded the capacity of these call centers. A news report by Sedensky [116] highlighted that the Houston 911 call center received and processed 75,000 calls in one day from August 28th to 29th, which was more than eight times its normal daily volume; this number does not account for the thousands of calls that were abandoned due to excessive waiting times. To address this situation, people turned to social media to share rescue requests and calls for help. Disaster relief organizations, assisted by many digital volunteers, sifted through social media to locate posts calling for help. For example, the Digital Humanitarian Network (DHN) sent USCG command centers information every six hours in Excel spreadsheets with the collected information needed for rescue. Two USCG Academy cadets organized 500 volunteers worldwide to read social media posts, and then they used GIS to create maps for a command center. The volunteers used hashtag and keyword searches to filter through millions of posts [134]. This underscores the importance of real-time rescue information collected through secondary sources, such as social media, for enhancing response processes. This research contributes to the fields of disaster informatics and disaster management by introducing novel methods for efficiently extracting rescue information on social media to assist decision-makers during hurricanes. It builds upon existing research in the literature, aiming to better integrate social media channels as a complementary yet relevant source of information during hurricanes.

1.5 DISSERTATION ORGANIZATION

Chapter 2 provides an overview of the related work. Chapter 3 describes the research methodology used in this research and introduces the proposed research methods and models. Chapter 4 presents the results of this dissertation. Finally, chapter 5 presents the main findings, limitations, future work, and implications of this research.

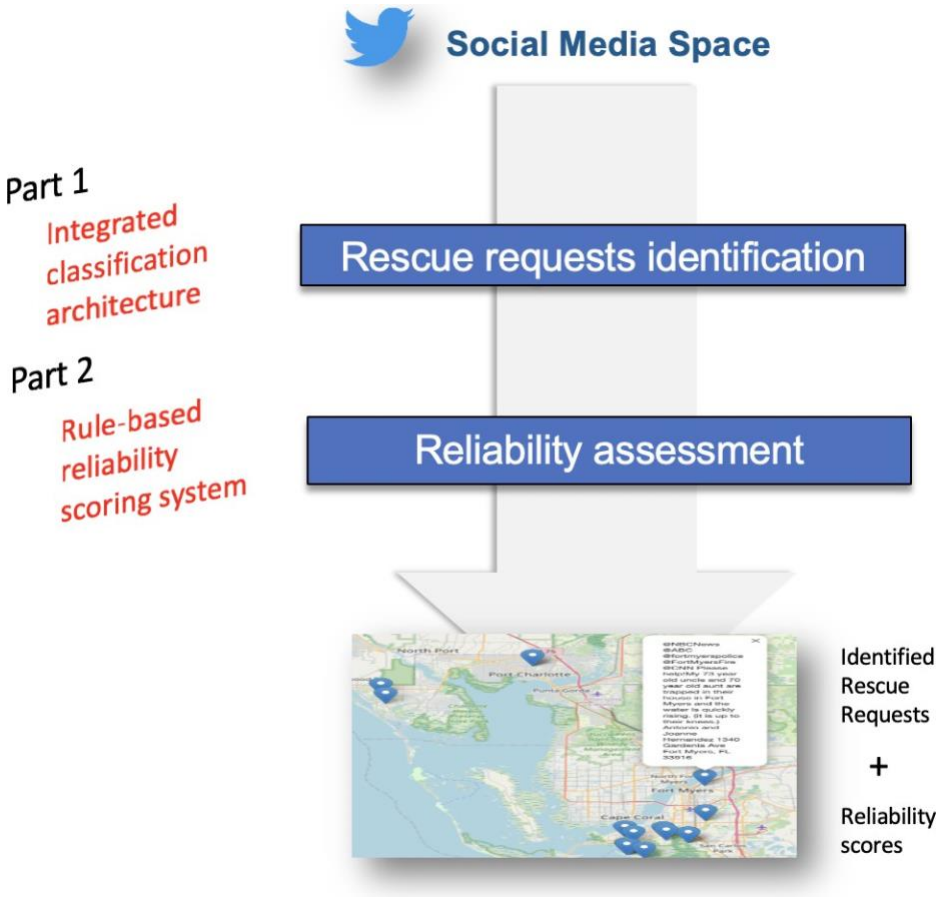

Figure 1. Proposed rescue identification framework

CHAPTER 2

LITERATURE REVIEW

Social media has recently become an integral component of emergency management and response. Traditionally, emergency response teams and government officials have used social media platforms as one-way channels to disseminate critical information, promote awareness, and orchestrate response strategies. Nonetheless, the availability of information technology devices, such as smartphones and laptops, and easy access to the internet access, has transformed social media into a two-way communication channel. Reports on Hurricane Harvey show that people used Twitter as an alternative for seeking/offering help and reporting situational information [18]. The large flow of messages posted on social media by the general public during natural disasters provides a valuable source of information for emergency responders to build situational awareness.

2.1 BACKGROUND INFORMATION ON DISASTER MANAGEMENT

The frequency and severity of natural disasters have increased, significantly impacting the lives of millions of people globally each year [81]. In the last decade, disaster events such as tsunamis, floods, earthquakes, and pandemics have affected over 2.6 billion people [26].

The emergency management process consists of four major phases: preparedness, mitigation, response, and recovery. This framework is proposed by the Federal Emergency Management Agency (FEMA) [40]. Understanding the activities associated with each of these phases is crucial. The preparedness and mitigation phases occur before a natural disaster, involving risk

assessment, resource, and expertise identification, and planning to minimize potential disaster impacts (e.g., developing evacuation plans). The response and recovery phases take place after a disaster and can include multiple activities such as the implementation of emergency plans, medical care, rescue activities, shelter management, distribution of supplies, damage assessment and prevention, and recovery efforts. These phases are illustrated in Figure 2.

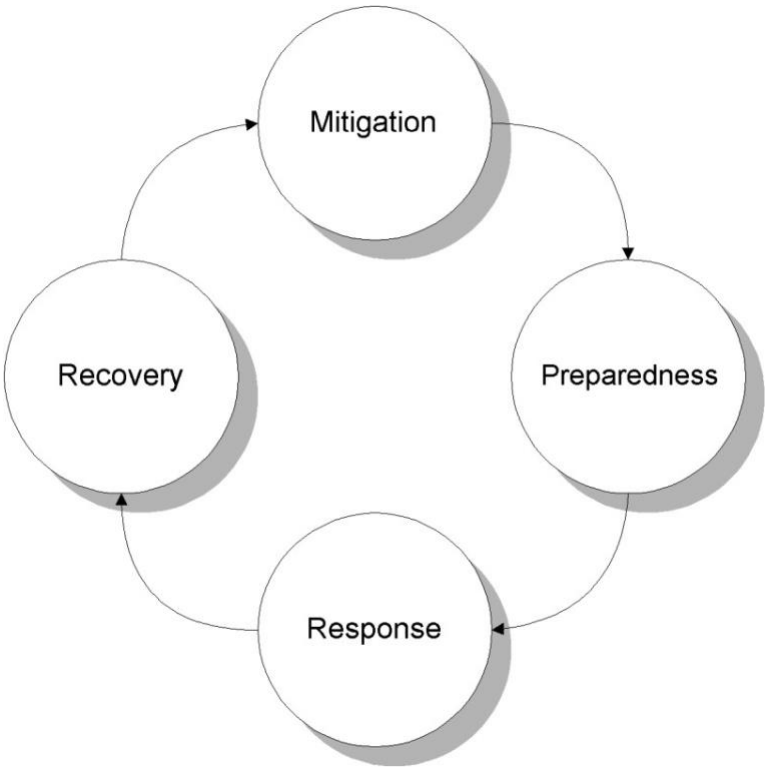

Figure 2. Emergency management phases according to FEMA

Disaster response is one of the most researched phases, as it is during this phase that immediate assistance becomes critical for affected people [81]. During this phase, the goal is to respond to the immediate needs and take actions that reduce fatalities, injuries, and damage loss. A rapid response can significantly save lives, meet the victim's needs, and mitigate the catastrophic impact

of the disaster. Disaster response involves multiple actors operating in an uncertain and complex environment, including local authorities, humanitarian organizations, and volunteers. In this environment, if the decision-making process—from data collection to decision-making—is too long, it may result in ineffective decisions based on an outdated operational picture [75]. Hence, processing real-time data is the key to effective disaster response decision-making. Information needed during the different emergency management phases can be broadly categorized into two classes of information:

- High-level information, such as the disaster size and magnitude, death rates, and affected areas, is typically used by top-level management for strategic decisions.
- Actionable (tactical) information, such as help requests, road closures, road flooding, and missing individuals, is low-level information types used for operational and tactical decisions.

Social media platforms provide a valuable source of real-time information during emergencies. Numerous studies have explored the increasing role of social media in disaster communication. For instance, Lachlan et al. [76] analyzed the common search strategies for disaster information from social media. They suggested using localized hashtags as an effective way to disseminate useful information. Other studies, such as those in [80] and [37], have tried to bridge the gap between research and practice by providing a summary of the most effective practices and lessons from using social media in disaster response. They noted that social media platforms, such as Twitter, are still utilized primarily as one-way communication channels; researchers have emphasized the importance of monitoring public feeds on social media to enhance the current emergency response

practice. Researchers in disaster informatics have made significant advancements in developing automated tools for extracting relevant information from social media platforms, employing methods from machine learning (ML), deep learning (DL), and artificial intelligence (AI).

2.2 DISASTER-RELATED CLASSIFICATION TASKS

This section explores automated techniques for extracting information through machine learning (ML) and deep learning (DL) methods. Previous research studies have formulated the problem as a classification problem, in which social media messages are categorized into several information types. Classification methods can be categorized according to their objectives into several tasks:

- Categorizing social media messages based on their informativeness (e.g., informative tweets and non-informative tweets)
- Categorizing tweets based on several humanitarian themes, including infrastructure damage, donation efforts, volunteering requests, and others.
- Identifying tweets posted by eyewitnesses in the disaster area
- Categorizing tweets based on the severity of the damages

Table 1 provides examples of research studies proposing automated methods for disaster-related classification, organized by the tasks they aim to accomplish.

Several public datasets have been used in prior research to train and evaluate the proposed classification methods, for instance, the disasterLexT26 dataset [105], disasterMMD [7], and HumAID [9]. These datasets comprise tweets collected through various disaster events and categorized by several information types. For instance, the disasterLexT26 dataset was created

Table 1. Related classification tasks in the literature

Category	Main labels	Examples
Classify by informativeness	relevant and non-relevant on-topic and off-topic informative and non-informative disaster-related and non-disaster related situational and non-situational	[107] [22] [77] [66] [13] [87] [70]
Classify by humanitarian categories	casualties and public impact missing trapped or found people displaced people and evacuations injured or dead people caution and advice sympathy and emotion infrastructure and utility damages	[141] [56] [4]
Classify by the source of the messages	direct eyewitnesses indirect eyewitnesses vulnerable eyewitnesses	[146] [130] [72]
Classify by the urgency of the messages	urgent rescue request	[34] [148] [137] [60]
Classify by damage assessment	severe damage mild damage low damage	[57] [8] [97]

by Olteanu et al. [105] in 2015 and was made publicly available through the disasterLex project[1]. This collection of disaster datasets was compiled from 26 crises between 2012 and 2013 and annotated by crowdsourced workers. These datasets were annotated for informativeness (informative vs. non-informative tweets) and humanitarian needs categories (e.g., affected individuals, infrastructure and utility damages, donation efforts and volunteering, caution and advice). The disasterMMD dataset was introduced by Alam [7] and published through the disasterNLP resources[2]. This dataset contains labeled information derived from the tweets' textual content and images. The datasets were annotated based on their informativeness and humanitarian needs categories. HumAID (Human-Annotated Disaster Incidents Data) [9] is one of the largest publicly available human-annotated Twitter datasets, consisting of 77,000 labeled tweets. These tweets were sampled from 24 million tweets gathered during 19 major real-world disasters between 2016 and 2019. The HumAID dataset includes tweets categorized into several humanitarian needs categories. In these datasets, tweets were annotated in two primary ways: firstly, at a high level to determine if a tweet is disaster-related or informative, and secondly, based on specific humanitarian needs. The categories for humanitarian needs include several classes, such as caution and advice, sympathy and support, requests and urgent needs, displacement and evacuation, injuries and fatalities, missing and found individuals, infrastructure and utility damage, as well as rescue, volunteering, and donation efforts. Most ML/DL systems developed in this area have focused on improving the accuracy of the proposed models, using these public benchmarks for evaluation. Consequently, researchers have defined various classification tasks derived from the public benchmarks' labels. These tasks include identifying

[1] https://disasterlex.org/
[2] https://disasternlp.qcri.org/

informative and disaster-related social media messages, categorizing social media messages by humanitarian information types (e.g., cautions and advice, sympathy and emotion, infrastructure and utility damage, etc), identifying social media messages posted by eyewitnesses, and extracting social media messages reporting damages, among other.

Categorization by informativeness – One straightforward application of machine learning in emergency response involves developing automated systems that can assist emergency response teams by identifying informative (sometimes referred to as disaster-related or on-topic tweets) social media messages from the huge volume of data posted on social media during disasters. Early studies that explored ML/Dl techniques for disaster response primarily focused on this task. For example, Parilla-Ferrer and colleagues [107] proposed a machine learning approach to identify informative tweets using custom data collected during the 2012 Luzon southwest monsoon floods. Caragea et al. [22] proposed a deep learning-based approach to identify informative tweets, evaluating their model using the disasterLexT26 datasets [105]. Li et al. [77] introduced a domain adaptation approach to classify tweets into on-topic and off-topic categories using disasterLex datasets. Alshehri et al. [13] developed an ensemble learning algorithm for the same task. Another line of research has focused on identifying informative disaster-related tweets by integrating multiple data modes, including both textual data and images. For instance, Koshy et al. [66], Kumar et al. [70], and Madichetty et al. [87] proposed multimodal deep learning models to filter informative tweets during natural disasters, using public benchmark datasets such as disasterMMD data set [7] for evaluation. Informative tweets typically refer to those social media messages that provide useful information directly related to the disaster. The primary issue with this classification is its vagueness; researchers have not provided clear guidelines on what

qualifies as a disaster-related or informative message. However, given the large volume of posts that decision-makers must process, this classification scheme is useful for high-level filtering social media messages.

Classification into humanitarian information types – Numerous research studies have suggested using machine learning and deep learning-based methods to classify tweets into various humanitarian categories rather than informativeness. To this end, researchers have defined multiple information types, and the problem has been formulated as either a multi-class or multi-label classification problem. For instance, Liu et al. [83] introduced a transformer-based technique that accomplishes three tasks: (1) identifying disaster-related tweets, (2) classifying tweets into various information types, and (3) recognizing rumor tweets. The authors evaluated their model using labels from the disasterLex T6 and T26 datasets. Similarly, Yu et al. [141] proposed a deep learning-based approach for cross-event topic classification, employing a classification scheme that included five humanitarian categories, namely caution and advice, casualties and damage, information sources, donation and aid, and missing people. A deep learning approach for multi- class classification was proposed by Aipe et al. [4], which was evaluated using the disasterNLP datasets. Instead of employing the disasterNLP humanitarian categories, the authors developed their classification scheme, consisting of the following categories: (1) Casualties and Public Impact, encompassing labels such as Injured or Dead People, Missing, Trapped, or Found People, and Displaced People and Evacuations from the original datasets; (2) Collateral Damages Class, derived from Infrastructure and Utility Damages in the original datasets; (3) General Awareness, created from the Caution and Advice class of the original dataset; (4) Voluntary Services Class, derived from the Donation class in the original

datasets; (5) Sympathy and Emotion, derived from the Emotional Support class; (6) Disaster-Specific Information, derived from the Other Useful Information class in the original dataset; and (7) Non-Informative Tweets.

Eyewitness identification task – Obtaining firsthand information directly from the disaster area is crucial for disaster response officials when a disaster occurs. Research has shown that local citizens and eyewitnesses are significant sources of information, as they share a vast amount of data. To identify such information, Zahra et al. [146] developed a method for identifying the direct and indirect eyewitness messages posted on social media during natural disasters. They created a classification scheme that categorizes tweets as 'direct', 'indirect', and 'vulnerable' eyewitness messages. Tanev et al. [130] proposed a learning framework focused on detecting micro-events (e.g., casualties, destruction, and damages) caused by natural disasters and identifying eyewitness reports in both types of media. Kumar et al. [72] proposed a multi-channel CNN deep learning framework for this task. Their model outperformed multiple conventional and deep learning models in identifying eyewitness posts.

Damage assessment task – Assessing the impact of a disaster, specifically infrastructure damage, is a crucial task in disaster response. Typically, disaster response officials rely on field experts to comprehend the extent of the damages and assess losses. Nonetheless, this approach can be time-consuming [57]. Therefore, numerous research studies have developed automated techniques for damage assessment. A common method of assessing damage through social media involves classifying tweets based on the severity of the damage, which typically falls into one of three categories: severe, mild, or low. Alam et al. [8] proposed a social media image processing pipeline to filter social media imagery content (e.g., removing duplicate images and retaining relevant ones) and extract actionable information about the damage severity. Nguyen et al. [97]

proposed a convolutional neural network (CNN) model to classify social media images into these three damage severity classes. They conducted an extensive evaluation, demonstrating the advantages of using CNNs over state-of-the-art machine learning models for this specific problem.

Previous research papers have relied solely on images to detect damages. A few studies have shifted their focus to identifying damage-related information within the textual content posted on social media. For instance, Madichetty et al. [88] proposed an ensemble-learning approach to detect tweets containing information related to damage assessment. Specifically, their approach targeted two types of damage: (1) infrastructure damage inflicted on specific resources such as roads, railways, and towers; and (2) human damages referring to tweets reporting injuries and deaths. In another work, Madichetty et al. [90] proposed support vector regression and random forest approaches to automatically identify tweets for damage assessment. They introduced several low-level lexical and syntactic features related to damage assessment. The majority of classifiers proposed multimodal approaches for damage assessment using both text and image modalities. Examples include the studies by Hao et al. [49], Abavisani et al. [1], and Gautam et al. [43].

Other disaster classification tasks – While most past research on social media disaster classification has focused on the tasks described above, it is worth noting that disaster classification research has addressed a broader range of tasks, and additional variants of the problem have been explored. Related problems include disaster text summarization, location identification, rescue requests identification, and others. Devaraj et al. [34], Zhou et al. [148], and Wang et al. [137] tackled the problem of identifying emergency rescue messages from social

media by framing it as a binary classification problem. Interpretable machine learning models through text summarization is an important and promising research direction to explore. Existing studies tend to rely on black-box machine learning approaches, which are not ideal for disaster response officials who require a clear understanding of the situation. Nguyen et al. [98] investigated the task of summarizing informative tweets and proposed a new approach that categorizes information into various humanitarian categories, followed by summarizing the information from each category. The authors developed a unique classification task that involves classifying rationales into different information types. 'Rationales' (also referred to as 'explanations') are short snippets in tweets that provide enough evidence to classify the tweet into a specific information type. In the tweet example cited by the authors [98], '03 Dec 2012 – At least 475 people are killed after Typhoon Bopha makes landfall in the Philippines', the snippet 'At least 475 people are killed' provides sufficient information to classify the tweet into the injuries and deaths category. The proposed framework automatically categorizes tweets once snippets are detected. The snippet provides interpretable information that can be summarized.

Identifying locations of people in need during natural disasters is another task that has attracted the attention of several researchers. While several studies have explored the possibility of using geocoordinate data from Twitter to obtain accurate locations, such tweets constitute only a small proportion of the total number of tweets posted in real-time. Kumar et al. [71] highlighted that only 7.9% of tweets are geotagged. Therefore, relying on the tweets' textual content to determine individuals' locations is a viable strategy to overcome this limitation. Karam et al. [62] presented a machine learning solution that uses Support Vector Machine (SVM) and Named Entity Recognition (NER) to infer the locations of those in need of help within the disaster area. Unlike

previous approaches that heavily relied on Named Entity Recognition (NER) and geotagging, Kumar et al. [69] introduced a CNN-based classification method to extract location information from tweets.

2.3 PROPOSED METHODS FOR DISASTER-RELATED CLASSIFICATION TASKS

This section reviews the methods proposed for the tasks presented above. The proposed methods can be broadly categorized into the following classes: (1) keyword-matching approaches, (2) ML methods, (2) DL methods, and (4) transformer-based methods.

2.3.1 Keyword-Based Methods

Keyword-matching methods have been employed as rapid and efficient techniques to process the vast volume of messages shared on social media platforms, such as Twitter, during emergencies. Keyword-matching generally involves selecting a set of predefined keywords or hashtags by users to identify relevant tweets.

Twitter APIs—search and streaming APIs—allow users to specify their queries to retrieve tweets based on specific criteria such as locations, search keywords, or hashtags. Hence, the Twitter API has been integrated into many social media analytics platforms for disaster management, such as AIDR [55] and Tweedr [15] platforms. The Twitter API relies on users specifying the right keywords for search based on their knowledge and expertise. Consequently, several research studies have explored how to develop more efficient keyword-matching approaches. Olteanu et al. [104] developed a more efficient keyword-based approach for querying relevant results from the Twitter API. This method starts by categorizing a sample of tweets as informative and non-informative, then assesses the terms from each category using statistical tests

such as Chi-square and PMI. The highest-scoring terms are then refined and used to construct a disaster- related term lexicon named disasterLex. Zheng et al. [147] introduced a semi-supervised dynamic keyword generation technique utilizing incremental clustering, support vector machines (SVMs), expectation maximization, and word graph generation to produce pertinent disaster keywords. This method leverages a limited number of labeled tweets and the co-occurrence properties of word to automatically generate and expand a list of keywords over time, capturing the evolution of the event. Olteanu et al. [104] introduced an advanced keyword-based method for extracting relevant results from the Twitter API. Their technique initiates by classifying a subset of tweets into informative (positive) and non-informative categories. Subsequently, terms within these categories are evaluated using statistical tests such as Chi-square and PMI, allowing for the refinement and selection of the most significant terms. These terms are then utilized to construct a disaster-related term lexicon known as disasterLex. Similarly, Zheng et al. [147] developed a semi-supervised dynamic keyword generation strategy that incorporates incremental clustering, support vector machines (SVMs), expectation maximization, and word graph generation to identify critical disaster keywords. This strategy employs a small set of labeled tweets and analyzes word co-occurrence to dynamically generate and update a keyword list that adapts to the unfolding of the event. Kumar et al. [73] developed TweetTracker, a platform that uses user-generated keywords to collect and monitor disaster tweets in real time, visualizing key keywords and hashtag trends. Hashtags have increasingly been used to filter relevant disaster tweets. Lachlan et al. [76] analyzed tweets from a winter storm and observed that localized hashtags were more likely to be linked to disaster-relevant tweets. They argued that utilizing local hashtags during a disaster event is an effective method for detecting actionable information. Hien To et al. [132] introduced a systematic approach to generating a comprehensive list of

relevant hashtags from a small set of keywords, which aids in identifying informative tweets from the vast data on social media.

Overall, keywords and hashtags-based approaches for identifying relevant disaster tweets typically depend on a small set of predefined keywords or hashtags, such as combinations of disaster names or types with names of affected areas, often resulting in missing relevant tweets. These methods are predominantly manual, making them unscalable and ineffective for real-time disaster applications. Given the vast amount of data available in real-time as a natural disaster unfolds, monitoring social media platforms by keywords is impractical. Hence, researchers have explored automatic methods to identify relevant information from social media.

2.3.2 Traditional Machine Learning Methods

Early research in this field employed traditional machine-learning methods to classify emergency- related tweets. These works used various machine learning models, mostly Support Vector Ma- chines (SVM), Support Vector Regression (SVR), Random Forest (RF), AdaBoost (ADA), Naive Bayes (NB), Linear Regression (LR), Decision Tree (DT), and AdaBoost (ADA). Nazer et al. [95] compared several models, including ADA, SVM, RF, and DT, to identify tweets specifying urgent needs (e.g., food, water, etc.). A data set of 3,261 help requests and 9,999 regular tweets collected from Hurricane Sandy was used to train the models. In this study, DT was the best-performing model (in terms of F1 measurement) Habdank [45] utilized custom Twitter data generated from an incident in Ludwigshafen, Germany, in October 2016 to evaluate SVM and RF on the relevancy assessment task (i.e., categorize tweets into relevant and non-relevant tweets). They trained several ML classifiers, such as NB, DT, RF, and SVM. The performance of RF was the highest across all measurements (accuracy, F1, recall, and precision)

among the tested classifiers. Parilla et al. [107] trained two machine learning models (SVM and NB) to identify disaster-related tweets using manually annotated data collected from the Manila flooding. In this research, SVM significantly outperformed NB in several metrics. Wang et al. [137] and Devaraj et al. [34] assessed several machine learning models for identifying urgent rescue tweets, including RF, SVM, NB, and others. Wang et al. [137] compared the performance of several traditional machine learning classifiers, such as RF, NB, SVM, and LR, and found that the SVM model outperforms the competing machine learning models on the rescue-seeking messages identification task. Madichetty et al. [90] proposed different types of features, including lexical, syntactic, and word frequency features, for identifying damage-related tweets. A novel two-stage machine learning approach was used. The first stage of training involved using SVR and LR to give weights to the different feature vectors. The weighted feature vector is used as an input to the RF classifier in the second stage. The authors performed extensive evaluations of their proposed method, utilizing multiple public benchmarks for disaster response, and demonstrated its effectiveness in diverse domains and disaster types.

2.3.3 Deep Learning-Based Methods

Various deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have been extensively employed to tackle the challenge of classifying social media messages during emergencies.

Several studies, including those by Yu et al. [141] and Caragea et al. [22], have explored the effectiveness of Convolutional Neural Networks (CNNs) in categorizing tweets into various humanitarian categories and identifying informative and disaster-related tweets. For both tasks, CNNs have outperformed the traditional machine learning models, such as SVM and LR. The CNN

provided better generalization across various events, unlike non-neural network models, which are domain dependent. As a result, when obtaining labeled data for new disaster events is difficult, CNNs were considered a better approach than traditional ML models.

Madichetty et al. [86] trained different deep learning models to detect situational tweets from social media, including CNN, LSTM, Bi-directional LSTM, and Bi-directional LSTM with attention (BLSTM-attention). Several word embedding models were evaluated in this paper, including those pre-trained on disaster data and those pre-trained for general use, and it has been found that disaster-related word embeddings performed better with the deep learning models. Among the different architectures tested in this paper, BLSTM with an attention mechanism showed a better performance for this task. Ning et al. [99] developed an automated method to identify and summarize informative tweets related to specific disaster events. Their model utilized a correlative CNN architecture that was trained to categorize tweets by informativeness and source. This architecture included a shared CNN module that transforms the input feature vector into a higher-level shared representation layer. This layer is then fed into two separate modules, each consisting of fully connected layers dedicated to one of the classification tasks. Additionally, the model incorporated a set of hand-crafted features, including lexical, emotional/sentiment, POS tagging-based features, and topical features such as n-grams and LDA topics. They found that the topical and linguistic features significantly enhanced the performance of the CNN model in both classification tasks.

2.3.4 Transformer-Based Methods

Recently, the application of transformer-based models for various disaster classification tasks has gained increasing interest. Liu et al. [83] developed disasterBERT, a transformer-based method

that was customized for disaster detection and recognition tasks. They trained a set of transformers to categorize tweets into binary categories (disaster-related and non-disaster-related) and multi-class categories (humanitarian information types). The authors conducted a large-scale evaluation of several transformer models, including BERT, XLNet, GPT-2, and RoBERTa, with DistilERT emerging as the most effective model. The proposed disasterBERT architecture comprises three layers: tokenization, transformation, and classification. Initially, tweets are tokenized using BERT's internal mechanism, incorporating special tokens such as 'CLS.' The tokenized outputs are processed through DistilBERT layers, followed by a linear classifier that leverages the 'CLS' embedding vector to categorize the tweets. The entire system is trained end-to-end, showcasing superior performance over traditional classifiers such as LR, NB, and SVM, and even outperforming deep learning models such as CNN and LSTM. Additionally, the authors introduced disaster2vec, a contextual word embedding model developed by pre-training DistilBERT on a substantial corpus of disaster-related data. This model is designed to enhance disaster classification tasks, offering an alternative to general-purpose word embedding models such as Word2vec [93]. Nguyen et al. [98] developed a two-stage BERT model (BERT2BERT) classifier within their text classification and summarization framework. In the first stage, they extracted rationales from tweets by jointly training two BERT models in a multi-task process. The first BERT model was responsible for categorizing the tweet into humanitarian categories, while the second model identified which tokens contained useful information to be tagged as rationales, assigning each token a binary label of either 1 or 0 to indicate its inclusion in the rationale information. In the second stage, the classification task from the first stage was disregarded, and a different BERT-based classifier was trained on the rationale vectors (stage 1 outputs) to determine the final classification results. Evaluated on two disasterNLP datasets for

humanitarian tweet classification, the proposed classifier demonstrated superior performance compared to various traditional machine learning classifiers. Li et al. [77] proposed a self-training domain adaptation approach for disaster tweet identification. This method involves pre-training a BERT model on a large set of unlabeled disaster-related tweets from previous disaster occurrences, followed by using the pre-trained BERT as a backbone model to transform input tweets into BERT embedding space. A CNN is then placed on top of the BERT model and trained to classify messages from a target disaster event. The evaluation results using disasterNLP and disasterLex datasets showed that the proposed self-training mechanism enhances the classification performance of CNN. It is recommended to employ self-trained transformers to transfer knowledge from previous events to a target disaster event. Furthermore, the study revealed that pre-training BERT on prior disaster-related tweets yields better results than using the standard BERT model pre-trained on the Wikipedia Corpus. Madichetty et al. [89] have recently presented a novel approach that combines a fine-tuned RoBERTa transformer with a feature-based method to classify situational tweets related to natural disasters. In their proposed approach, two feature vectors are generated: the first is created by training an SVM on lexical and syntactic features extracted from tweets, such as subjective words, personal pronouns, numeral counts, exclamation and question marks, and slang. The second feature vector is obtained by fine-tuning a RoBERTa transformer. The authors employed a multiplicative fusion technique (element-wise multiplication) to merge the two feature vectors into a single vector, which is then used for the final classification. Compared to several deep learning models, including CNN, LSTM, and LSTM with an attention layer, their approach has demonstrated superior performance.

2.4 EMERGENCY RESCUE REQUESTS IDENTIFICATION PROBLEM

2.4.1 Social Media Rescue Messages Characteristics

Understanding the characteristics of rescue messages posted on social media is an important step toward designing robust classifiers to automatically identify these messages in real-life scenarios. However, it is worth noting that only a few research studies have examined their characteristics. The purpose of this review is to highlight the key features of rescue request messages as identified in the related literature. some of these features are included in this dissertation, such as:

- Rescue hashtags
- Mentions of emergency response organizations' accounts
- Location information
- Situational description

Numerous studies have examined the use of rescue hashtags as a means of assisting victims of natural disasters. During Hurricane Harvey, hashtags such as #SOSHarvey, #SOSHouston, and #Rescue were employed to highlight social media messages from those in immediate need of assistance. However, many studies have noted that most rescue tweets do not contain any rescue hashtags. Zou et al. [149] observed that rescue hashtags are used in only a small fraction of tweets. Their textual analysis of an annotated sample of tweets revealed that the most frequently used rescue hashtags during Harvey were #harveysos, #houstonflood, #rescue, #houston, and #houston-strong. These were categorized as either location-based (local) hashtags or rescue-based hashtags that include rescue and help terms [149]. Based on an online survey, Mihunov et al. [92] found that about 25 percent of rescue tweets during Hurricane Harvey utilized rescue hashtags. The event

was marked by a variety of hashtags, such as #HurricaneHarvey, which were not necessarily used for rescue requests. Lachlan et al. [76] investigated how hashtags have been used to find actionable social media messages during natural disasters. They found that local hashtags were more effective for filtering rescue requests than nationwide hashtags. In general, relying solely on rescue hashtags is insufficient to identify rescue tweets. Nonetheless, they can still serve as a useful feature in defining rescue messages. Zou et al. [149] reported that it is a common practice to tag emergency response organizations and government officials (using the mention feature '@') for seeking help. Houston police and Cajun Navy (informal volunteer groups) accounts were often tagged during Harvey. However, about 45.84% of the analyzed rescue tweets did not contain this feature. Song et al. [127] analyzed a dataset of verified rescue request tweets from the 2018 torrential rains in Western Japan in order to extract useful rescue features that could be employed in automatic classification methods. Their analysis identified several types of information in rescue requests, including (1) rescue details, (2) location information, and (3) rescue hashtags. They categorized the identified features into those that can be used to detect rescue requests and those that cannot. They suggested that both location and rescue details features are essential characteristics of rescue tweets. Rescue details may include details related to the disaster situation, such as descriptions of evacuation locations, damage reports, or reports of evacuation difficulties. They may also include clear references to the relationship between the victim and the eyewitness reporting the message (e.g., whether the post is on behalf of a relative or a friend). For instance, phrases like 'my father' and 'my friends' were commonly used in rescue request messages. Location feature specifies the location of the person who needs rescue, which may be conveyed through a full address, including house number, city, or building name, or through location-related hashtags.

Zou et al. [149] analyzed an annotated dataset of emergency rescue tweets collected during Hurricane Harvey. They found that 89.48% of the collected rescue tweets contained full addresses, while 66.02% described situational information (e.g., information about victims). Only 61.08% of the tweets included both details. The authors also examined the number of hashtags attached to the annotated rescue messages and found that the majority of rescue tweets do not include rescue hashtags.

2.4.2 Automatic Methods for Rescue Requests Detection From Social Media

Devaraj et al. [34] conducted a study in which they evaluated a set of classifiers, including Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), AdaBoost, Multi-layer Perceptron (MLP), Logistic Regression (LR), and Convolutional Neural Network (CNN), for the task of detecting rescue messages on Twitter. The authors found that the average word embedding (i.e., creating a fixed-length representation of an input tweet by averaging the word embeddings of each word) is an effective approach for feature extraction. They also found that SVM and CNN were the most effective models for the problem measured in terms of F1 score. All the experiments of this study were performed on an annotated set of tweets collected during Hurricane Harvey.

Zhou et al. [148] introduced 'VictimFinder', a collection of transformer-based classifiers designed to identify rescue request messages from social media during natural disasters. To assess the effectiveness of the proposed models, the authors conducted experiments using a manually annotated set of tweets collected during Hurricane Harvey. They evaluated multiple transformer models such as BERT, ELMo, DistilBERT, ALBERT, and XLNet. The rescue message identification problem was divided into three distinct sub-tasks: (1) determining whether a tweet requests help, (2) determining whether it specifies a full address, and (3)

determining whether it provides detailed information about the victims. Each of these tasks was formulated as a binary classification problem. VictimFinder architecture consisted of a fine-tuned transformer with a classification head on top of it. Two training strategies were used: (1) fine-tuning strategy and (2) feature-based strategy. Fine-tuning involves training the backbone transformer models and updating their initial parameters using the custom training data from a target event. In contrast, the feature-based training strategy involves freezing the initial parameters of the backbone transformer, previously trained on large volumes of data, such as Wikipedia corpus, for general use. To evaluate the effectiveness of the proposed models, the authors performed experiments on a manually annotated set of tweets collected during Hurricane Harvey, using multiple transformer models such as BERT, ELMo, DistilBERT, ALBERT, and XLNet. The rescue messages identification problem was divided into three distinct sub-tasks: (1) determining whether a tweet requests help, (2) determining whether a full address is specified, and (3) determining whether detailed information about the victims is provided in the tweet. Each task was formulated as a binary classification problem. Two training strategies were used for VictimFinder: fine-tuning strategy and feature-based strategy. Fine-tuning involves training the backbone transformer models and updating their initial parameters using custom training data from a target event. In contrast, the feature-based strategy involves freezing the initial parameters of the backbone transformer, previously trained on large datasets such as the Wikipedia corpus, during the training stage. The study found that all BERT-based models outperformed conventional machine learning models, with BERT-LSTM and BERT-CNN models achieving the best performance. ELMo transformer achieved the highest recall but had slow data processing times. Although the ELMo transformer achieved the highest recall, it was slower in processing data. In their recent work, Wang et al. [137] presented a machine learning approach to detect

rescue-seeking messages on social media, which were strictly defined in this study as messages containing a precise address. The research explored how ZIP codes might in- fluence the performance of classifiers for rescue tweets. Several machine learning models were evaluated, including Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), and Logistic Regression (LR). The findings indicated that conventional machine learning models are more effective at detecting rescue tweets tagged with ZIP codes (i.e., victims posting rescue request tweets that include ZIP codes) in various scenarios, with RF yielding the best performance among all the models.

2.5 RELIABILITY ASSESSMENT OF SOCIAL MEDIA DATA

Assessing the reliability of online information is a longstanding problem that has attracted the interest of researchers. With the rise of social media platforms, this issue has become more challenging. Researchers have addressed various related problems, such as the identification of misinformation, rumors, and fake news. Related studies have explored the problem across different application domains, including general news, politics, healthcare, sports, and emergencies.

2.5.1 Related Problems

This section reviews the methods proposed for various related reliability assessment problems, such as fake news, rumor, and misinformation detection. An increasing number of people nowadays seek news from social media platforms rather than traditional news outlets. This shift is driven by the timely and cost-effective access to information that social media offers compared to traditional media. Consequently, large volumes of fake news (i.e., news containing

intentionally false information to deceive users) are produced and propagated through these platforms for various purposes, including political and financial gains [121]. This problem has been addressed extensively in the literature. Researchers have developed a wide variety of methods to automate this process.

State-of-the-art NLP models, such as BERT, have been employed for this problem and have shown promising results. Rai et al. [110] employed a BERT-LSTM model for fake news detection. The proposed model was trained to categorize news titles into fake and legitimate news and was effective when compared to several neural network methods. Kaliyar et al. [61] introduced FakeBERT, a BERT-based deep learning model, for detecting fake news. This model integrates several parallel CNN blocks and utilizes BERT as a sentence encoder for each incoming news report. Extensive evaluations conducted on a dataset from the 2016 U.S. presidential election demonstrated that FakeBERT outperforms traditional neural network models, such as CNN and LSTM. Choudhary et al. [27] presented a BERT-based deep learning framework (BerConvoNet) for classifying tweets into fake and real tweets. The proposed model includes two main components. A new embedding model using BERT for encoding input text and multi-scale feature block (MSFB) that includes multiple kernels to process text.

Silva et al. [122] proposed a multimodal approach for fake news detection. This study addresses the problem of poor generalization of ML/DL models across multiple news domains, such as healthcare and politics. The proposed model uses an unsupervised learning algorithm to learn embeddings from data across various domains and then employs a deep learning model that leverages this knowledge to perform domain-independent fake news classification. The proposed cross-domain classifier demonstrated promising performance, outperforming existing fake news

detection models in terms of F1 score, even with a low labeling budget. Yuan et al. [142] proposed a graph-attention neural network model to address this issue. This study also focused on the challenge of poor cross-domain generalization inherent in the proposed ML/DL methods. The model's architecture processes multimodal data. A BERT model converts input text, while a CNN variant, VGG-16, extracts visual features. Additionally, the model introduces a domain discriminator to achieve a robust representation for each domain by identifying domain-specific information. Kumari and Ekbal [74] proposed a multimodal deep learning architecture for fake news detection on social media by analyzing both textual and visual features. The proposed model utilizes an Attention-Based Stacked Bidirectional Long Short-Term Memory (ABS-BiLSTM) model to encode text and a Convolutional Neural Network–Recurrent Neural Network (ABM-CNN–RNN) for visual features extraction from the attached images. The extracted features are then integrated using Multimodal Factorized Bilinear Pooling (MFB), with a Multi-Layer Perceptron (MLP) per- forming the final classification. The proposed approach has been shown to outperform the state-of-the-art multimodal deep learning architectures. In addition, the authors found that the fusion of textual and visual data improved the performance of fake news detection classifiers. Li et al. [79] adopted a self-learning approach to enhance the classification accuracy of deep learning models for fake news detection. Their model automatically adds predicted samples with high confidence back into the training loop to improve the classifier's overall performance. Other research studies have developed automatic models specifically designed to detect misinformation and fake news during the COVID-19 outbreak. For instance, Kou et al. [67] introduced a knowledge-graph-based approach for detecting misinformation during the pandemic. This framework populates the graph network with pandemic-related knowledge facts crowdsourced from both experts and non-experts, using the knowledge graph to

provide an explainable outcome for users. Yue et al. [143] proposed a contrastive adaptation network for misinformation detection during the COVID-19 pandemic. This model employs a pseudo-labeling technique for low-cost data labeling and leverages a contrastive adaptation loss function to train the model. The objective is to make the model learn a robust domain-invariant embedding across several domains.

Researchers also focused on assessing the reliability of information sources rather than the input message itself. For instance, Jose et al. [59] investigated the problem of detecting spammers on social networks. They proposed a hybrid approach that combines k-means clustering and Latent Dirichlet Allocation (LDA) algorithms to group user-profiles and SVM to categorize users into spammer and non-spammer categories. Lu et al. [85] employed a graph neural network to predict whether a given tweet's source is fake. To this end, they introduced a new model called Graph-aware Co-Attention Network (GCAN), which provides an explainable classification outcome and a better representation of user interactions and the propagation of retweets. Scarlet et al. [112] proposed an attention-based graph neural network that predicts user actions and identifies social media accounts that are likely to spread and endorse misinformation on social media platforms. The framework, called SCARLET, models social media users as nodes in a graph network and is trained to identify vulnerable nodes based on historical behavioral data.

2.5.2 Reliability Assessment in Disaster Management Context

Most previous studies have narrowed down their research to a few application domains, such as general news, politics, and healthcare. Throughout this literature review, it has been noticed that less research has been conducted on assessing the reliability of social media and online data in the context of natural disasters and crises.

Krishnan et al. [68] trained an SVM classifier to identify tweets containing fake news during natural disasters. The authors utilized several user-related and content-related features to train the classifier, including the number of followers, the number of friends, the friends-to-followers ratio, sentiments, and URL reliability, among others. The proposed SVM classifier was evaluated on a set of tweets collected during Hurricane Sandy, the Paris attack, the Boston Marathon bombing, and other disaster events. The SVM demonstrated better performance compared to a decision tree classifier for disaster-related tweets. Hunt et al. [53] proposed a machine learning approach to assess the veracity of tweets posted during disaster events. The proposed approach categorizes tweets into three classes: (1) positive (i.e., the tweet contains reliable content), (2) negative (i.e., the tweet contains misinformation and false content), and (3) neutral (i.e., the tweet does not offer either positive or negative information). The authors annotated data collected from several disaster events. A comparative study was conducted to evaluate a set of ML algorithms, including k-nearest neighbors (KNN), decision tree (DT), random forest (RF), XGBoost (XGB), AdaBoost (AB), support vector machine (SVM), and multilayer perceptron (MLP). The results showed that SVM was the best-performing model among those evaluated. Rajdev et al. [111] developed a classification approach to detect malicious profiles during disaster events, with a case study focusing on the 2013 Moore Tornado and Hurricane Sandy. The authors employed a wide range of features derived from Twitter users' metadata, including both user-related and content-related attributes, to train the classifier. Yang et al. [139] proposed a transfer learning approach to detect misinformation on social media networks during disaster and disaster events. Pandey et al. [106] trained an ensemble learning classifier to identify reliable users on social media networks during disaster events and natural disasters. The proposed model automatically categorizes social media users into organization-affiliated and non-affiliated groups. To train the model, the authors

utilized various features derived from users' profile information, including social features (e.g., friends count, favorite counts) and activity features (e.g., status count, number of days since account creation). Halse et al. [46] investigated the impact of perceived emotions as predictors for social media messages' trustworthiness during disasters. Using datasets gathered during Hurricane Sandy in 2012 and the Boston Marathon bombing in 2013, the authors found that fear and neutral emotions significantly influence the perceived trustworthiness of a social media message.

Other research studies have conducted content analyses on datasets of tweets collected during disaster events, such as Hurricanes Sandy and Harvey. Hunt et al. [54] analyzed the spread of misinformation and rumors during these disasters through social media networks. This study closely examined case studies of rumors such as 'immigration status checks at shelters.' The authors categorized a set of tweets into five groups, including rumor-debunking, rumor-spreading, and rumor-questioning tweets, among others. They discovered that during these hurricanes, posts from verified accounts, such as those of government entities, received more interaction, which highlighted the importance of official accounts in rumor debunking. They also found that URLs linking to external sources, such as government agency websites, were frequently used as a de- bunking strategy by the public. This research emphasized the critical role of government agencies in combating misinformation by putting a system in place to use social media accounts as part of their effort to stop the misinformation spread. Wang et al. [136] analyzed a set of tweets posted during Hurricane Sandy, using a sample posted by official agents' accounts, including government organizations, NGOs, news agencies, and others. They identified the main characteristics of the collected accounts using several key metrics, such as

impressions, likes, mentions, retweets, and response time. The tweets posted by different profile categories were classified into various information types. The authors noted many significant differences among these accounts in terms of the number of retweets, interactions, and shared information.

Alrubaian et al. [11] proposed a scoring system to evaluate the reliability of messages posted on Twitter during disaster events. Their model comprises several components that analyze the tweet's content, user reputation, and user expertise, linked together in an algorithmic form. The proposed system employed several metrics to measure reliability for each component, including account popularity, users' sentiment history, and user activities, among others. The authors evaluated their model using real-world data collected during a disaster event (the Saudi-led campaign against Houthi rebels in Yemen). The proposed reliability scoring model demonstrated good performance in determining reliable and unreliable tweets from collected data. Assery et al. [16] proposed a semi-supervised model to evaluate the reliability of disaster-related tweets. This model also utilizes several features derived from user profile metadata and tweet content, employing a 10-point scale to assess the reliability of incoming tweets. The authors labeled two datasets from Hurricanes Michael and Florence. The evaluation results showed good accuracy of this model compared to a set of machine learning models trained for this purpose.

2.6 SUMMARY OF THE RESEARCH GAPS

A major research direction in disaster informatics literature has focused on exploring automated methods using ML and AI technologies to categorize disaster-related information. Researchers aimed to reduce the cognitive load on disaster responders by automating the process of filtering tweets, either based on their informativeness or by information type. They proposed several

classification tasks to achieve this goal.

Emergency rescue requests identification problem – As pointed out by Zade et al. [144], most of the previous classification approaches were trained to categorize tweets into general and broad information types, primarily organized around the idea of 'situational awareness' rather than the 'actionability' of the extracted information. This led to a low adaptability of these automated tools within formal disaster response workflow. Therefore, many research studies have suggested addressing the 'informational overload' problem on social media by designing methods guided by the concept of 'actionability'. Despite its practical relevance, only a few research studies have focused on extracting actionable information from social media. Although there is no definitive definition of actionable information on social media, most of the definitions given by disaster response practitioners (through previous interviews) consider actionable information as any piece of information (or request) that can be used to assist and respond to an identified issue. Such issues may include trapped victims, missing persons, road closures, and other similar emergencies. Therefore, more research is needed to address classification tasks for identifying actionable information. Among these tasks, identifying emergency rescue messages posted on social media platforms has not received much attention. The proposed studies focusing on various disaster-related information extraction tasks predominantly employ automatic learning methods. These methods employ traditional machine learning models, such as SVM, NB, and RF, and deep learning models, such as BERT, LSTM, and CNNs. The classification methods generally require a large number of labeled training samples to obtain strong classification performance. The labeling process is both time-consuming and resource- intensive. Furthermore, labeled training data may not be available in a short time as new disaster unfolds. In disaster scenarios, where

reactions vary significantly across different types of events, larger training samples are needed to address unseen disaster events. Although prior research re- ports good performance using supervised classification models for the different disaster classification tasks, there is a need to shift towards exploring semi-supervised, unsupervised, and few-shot learning approaches to address the data annotation challenge. Rule-based methods, employing domain-specific knowledge, often provide more precise results on small samples. However, since they do not require any training process, they fail to provide accurate results when evaluated on larger data sets. Researchers in different application domains, such as healthcare, have attempted to combine both approaches to obtain stronger classification outcomes. However, a combined learning and rule-based approach has not yet been utilized in disaster-related social media classification tasks, except in Zahera et al. study [145].

Reliability assessment problem – The reliability assessment of online data, including news web-sites and microblogs, has received significant attention across many domains, such as healthcare and politics, among others. The reliability assessment problem is very broad, covering numerous research tasks. Some of the major tasks that were extensively investigated include fake news detection, misinformation detection, rumors detection, and disinformation detection, among others. Researchers have distinguished between false information intentionally and non-unintentionally shared by users. Additionally, they investigated the problem of assessing the reliability of social media accounts (information sources) to determine spammers and bot accounts among regular accounts. The proposed studies have focused on a few application domains. For instance, the problem of misinformation detection in general news and during political events has been extensively investigated. During the COVID-19 pandemic, numerous

research studies explored automatic tools to detect fake news, which posed a significant challenge at the time. Healthcare is another application domain that has attracted considerable research attention. However, the reliability assessment problem of online data has drawn less attention in the disaster informatics domain. During disasters, certain types of information posted on social media platforms, such as calls for help, requests for assistance, help offers, road closures, and others, can be difficult to verify. While it is easier to create ground truth data for fake news detection using fact-checking platforms, information of this nature requires manual inspection and ground verification either by internal sources or through additional channels, such as emergency services 911. Such information types are very useful in times of disaster and can provide better insight; they can be used to take action directly or as part of building situational awareness (i.e., a global picture of the situation by decision-makers). Disaster response practitioners have highlighted that the verification of such information is of critical concern. There is a noticeable gap in the disaster informatics literature concerning the reliability assessment of actionable information posted by the public on social media. Developing automated tools to verify information, or at least to evaluate their reliability, would help decision-makers make better use of social media information. However, a significant challenge in this direction is the lack of annotated ground truth data. Explainability plays a crucial role in the reliability assessment tasks. Despite the success of existing machine learning models in many related problems, such as fake news detection, most of these methods are black-box methods that do not explain ‘why’ a piece of information is labeled ‘fake news’ or ‘misinformation’ [119]. Explainable outcomes can build trust in reliability assessment systems, leading to better decision-making. Hence, there are many ongoing research efforts to explore novel explainable methods for online data reliability assessment tasks. This dissertation uses a rule-based approach

to evaluate the reliability of rescue messages, employing a set of reliability indicators that are inherently explainable.

Most of the prior research has typically formulated the reliability assessment problem as a binary classification task. The proposed models categorize social media messages or online articles into either fake news (or misinformation, disinformation, rumor, etc) or not. Only a few studies have introduced scoring systems that assign a reliability score to a piece of information instead of a binary label. reliability assessment is a continuous process. Hence, there is a profound need to design a reliability assessment algorithm that performs a real-time assessment of the information. Modeling the reliability outcome as a continuous reliability score is a better approach to address this challenge.

CHAPTER 3

METHODOLOGY

This chapter discusses the methodology used in this dissertation and describes the proposed models. Section 3.1 briefly overviews the research methodologies employed for the rescue-seeking messages identification and reliability assessment problems. Section 3.2 describes the different hurricane events from which social media messages (tweets) were collected. Sections 3.3 and 3.4 present the different methods used for each problem.

3.1 BRIEF DESCRIPTION OF THE RESEARCH METHODOLOGY

This dissertation introduces novel models for identifying actionable rescue messages and assessing their reliability. This dissertation is grounded in a post-positivist, empiricist worldview [20] [36]. A quantitative research paradigm [32] was employed to evaluate the effectiveness of the proposed models. Historical tweets collected via Twitter APIs were used for empirical evaluation.

3.1.1 Research Design for the Rescue Messages Identification Problem.

The first part of the dissertation focuses on identifying rescue-seeking tweets. A quantitative research design was employed to answer the research questions related to this part. Figure 3 illustrates an overall description of the research design.

A set of features that characterize rescue tweets is proposed. These features are derived from existing literature and the author's analysis of several rescue-seeking messages posted during

various hurricanes. For example, people often include expressions for asking for help, such as ‘family trapped’ and ‘please help’ to request help. This study hypothesized that integrating domain-specific features (corresponding to specific textual patterns related to rescue messages) with large language models, such as BERT, would enhance the classification performance compared to the existing methods.

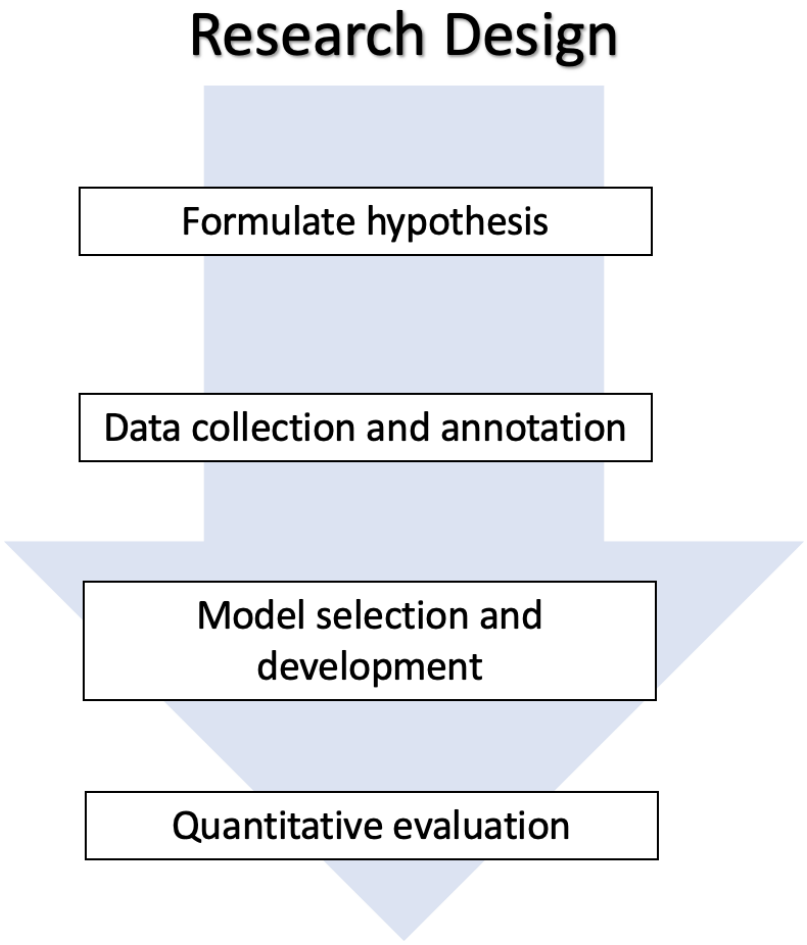

Figure 3. Design for the rescue tweets identification

The data for this dissertation were collected using Twitter APIs. Subsets of tweets were labeled as either emergency rescue tweets or non-emergency tweets using a predefined annotation

guideline. The labeled tweets were collected from various historical hurricane events, including Harvey (2017), Ida (2021), and Ian (2022).

The model's performance was evaluated using several statistical metrics, including F1 score, recall, accuracy, and AUC-PR. These metrics are commonly used in data science and machine learning studies to assess proposed models' performances. The proposed classification model in this study was compared to previous methods proposed in [148] and [34]. Given the relatively small size of the collected data, *k*-fold cross-validation was used for the comparative analysis. This method allows using all the annotated tweets at least once for evaluation. The predictions given by each classifier, including both accurate and misclassified tweets, were further analyzed to identify the strengths and the limitations of each method.

3.1.2 Research Design for the Reliability Assessment Problem

A quantitative research design was employed for the reliability assessment problem. Figure 4 illustrates an overall description of the research design.

Since no existing model or dataset can be directly applied to the reliability assessment of hurricane rescue messages, the related literature was reviewed to select a set of reliability indicators (factors) for building the proposed reliability assessment model. An informal discussion was held with an expert to refine the selection and gain more insights. The selected indicators were then quantified using metrics from previous research and existing datasets for similar problems. A rule-based model for reliability assessment was developed using these indicators. The identified rescue-seeking tweets from the previous part were used to evaluate the proposed reliability scoring model. A subset of rescue tweets was annotated by a predefined

reliability criterion. An annotation procedure was proposed to assign reliability labels to the rescue tweets using an external source provided by FEMA (FEMA damage assessment map). Since verifying the accuracy of rescue tweets post-disaster is challenging due to human and technical constraints, this study's proposed annotation scheme was designed to obtain an informed approximation of the rescue messages' trustworthiness. The output of the proposed scoring model was compared against the reliability labels assigned to the rescue tweets. The proposed reliability model was empirically compared to a set of machine learning algorithms and a competing method proposed by Assery et al. [16].

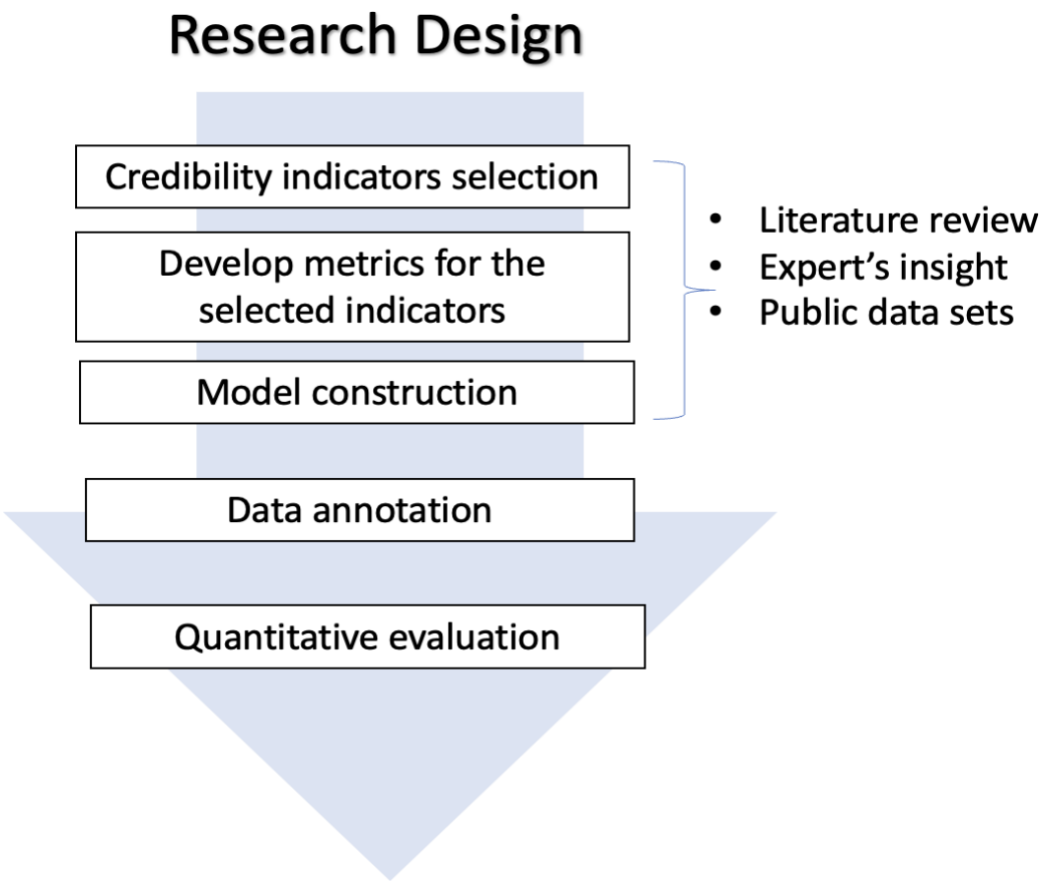

Figure 4. Design for the reliability assessment problem

The machine learning approach was selected for comparison because it is a common approach used for similar problems, such as misinformation, rumors, fake news, disinformation detection, and other related problems. The model by Assery et al. [16] was selected for comparison because it was designed to evaluate the reliability of tweets posted during disasters, even though it does not focus directly on rescue information. A 5-fold cross-validation approach was used in the evaluation of the machine learning models due to the small number of annotated rescue tweets.

3.2 HURRICANE EVENTS

Tweets posted during three hurricane events were collected: (1) Hurricane Harvey (2017), (2) Hurricane Ida (2021), and (3) Hurricane Ian (2022). Harvey was a disastrous Category 4 hurricane that made landfall in Texas and Louisiana on August 25, 2017. Harvey inflicted severe loss in terms of human lives (at least 88 deaths have been reported) along with substantial economic damage (133.8 billion) [133]. The damage was primarily caused by massive flooding that followed the hurricane, particularly in the Houston metropolitan area and southern Texas. The flooding displaced more than 30,000 people and resulted in more than 17,000 rescues [63]. Ida was another extremely destructive Category 4 hurricane that hit the southeastern region of Louisiana on August 29, 2021. The hurricane caused an estimated 75 billion in property and infrastructure damage [31]. As of September 9, 2021, 91 deaths have been reported across nine states. Most deaths occurred by drowning (60.4%), which was caused by the severe flooding that followed the hurricane [48]. Lastly, Ian hit the southwest coast of Florida on September 28, 2022, as a Category 4 hurricane. The hurricane caused more than 50.2 billion in terms of economic damage and over 150 direct and indirect deaths [82].

3.3 METHODS FOR EMERGENCY RESCUE REQUESTS IDENTIFICATION PROBLEM

This research presents a machine learning-based framework to automatically identify emergency rescue messages posted on Twitter during hurricanes. The proposed framework consists of three modules, as shown in Figure 5: (1) data collection and annotation, (2) classification, and (3) visualization. The data collection and annotation module are responsible for gathering and labeling data to train the classifier. The classification module is designed to automatically identify emergency rescue messages. To this end, a novel deep learning architecture that integrates low-level statistical features (extracted through a pre-trained BERT model) and high-level problem-specific features (derived from regex filters) was introduced. Finally, the visualization module is responsible for plotting the extracted rescue messages, thereby providing actionable information for emergency responders.

3.3.1 Problem Formulation

The rescue requests identification problem was formulated as a binary classification problem. Let $S = S_1, S_2, \ldots, S_n$ be a finite set of n tweets collected during a hurricane. Each tweet S_i in the sample is a sequence of T tokens $S_i = S^1, \ldots, S^T$. Let φ be the set of labels associated with these tweets. $\varphi \in 0, 1$ where $\varphi = 1$ indicates that the tweet contains an emergency rescue request, $\varphi = 0$ indicates otherwise. The classification problem can be formulated as a learning function f that maps the input tweets (transformed into a feature space) to the label space:

$$f : S \rightarrow \varphi$$

The objective is to train a classifier that minimizes the difference between the predicted labels and

the real labels based on a cost function *J*.

$$\min J(\varphi, f(S))$$

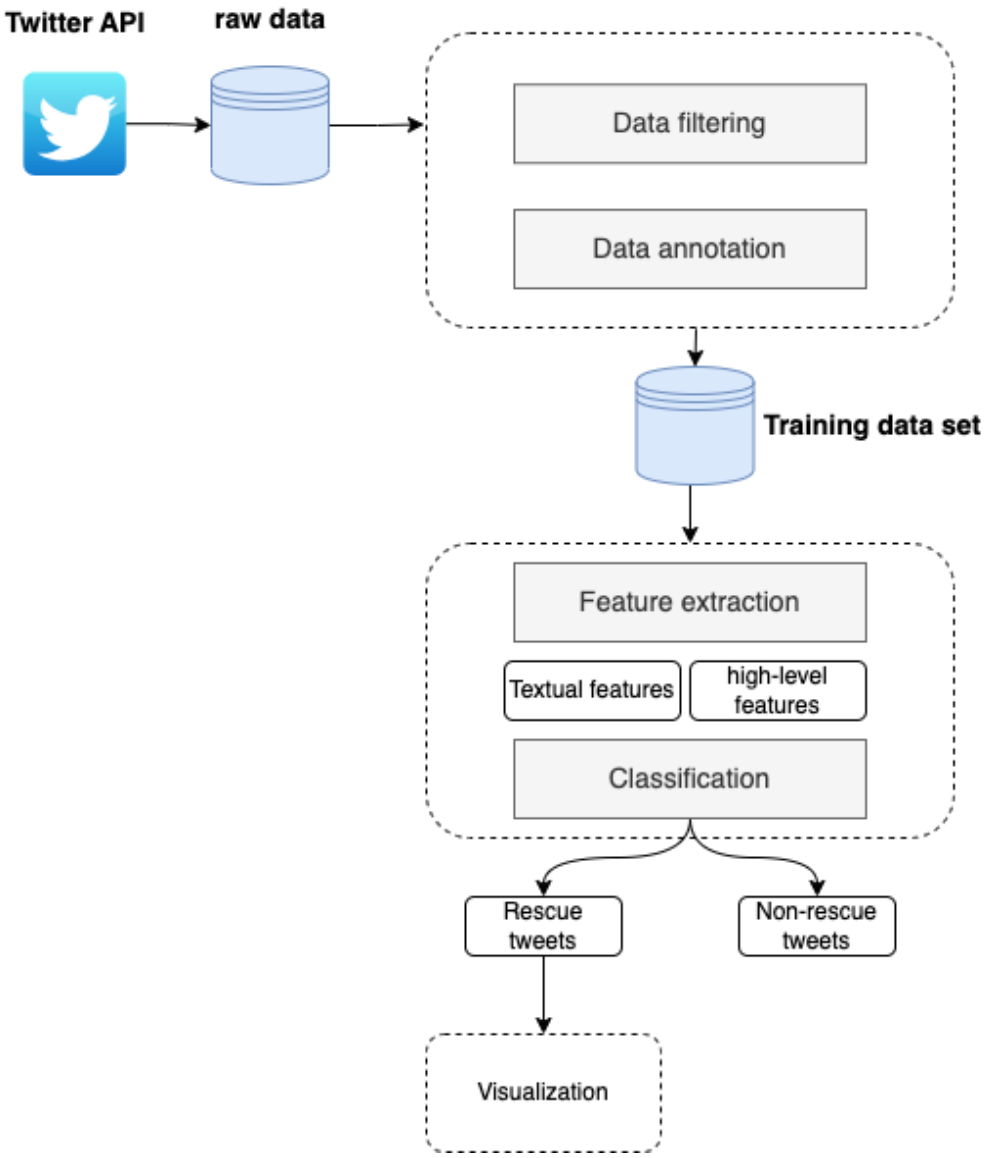

Figure 5. Emergency rescue requests identification framework

The most common cost function is the binary cross-entropy function (logistic loss):

$$L(y, \hat{y}) = -(y \cdot log(\hat{y}) + (1 - y) \cdot log(1 - \hat{y}))$$

where y and $\hat{y}$ are the true label and predicted label, respectively.

3.3.2 Data Collection

Due to the scarcity of publicly accessible data sets dedicated to this problem, the author of this dissertation collected and annotated a custom data set using Twitter streaming and search APIs.

Twitter provides a useful API for collecting real-time tweets. Tweets were collected from August 26 to August 31, 2017, during Hurricane Harvey. The Twitter streaming API allows the collection of tweets using several operators, including keywords, hashtags, and bounding box operators. Custom filtering rules can be created with several logical connectors, such as 'AND', 'OR', and 'XOR'. This study used hurricane-specific keywords such as 'Hurricane', '#Harvey', '#Hurricane- Harvey', and 'flooding' with the logical operator 'OR'. Furthermore, the longitude/latitude pairs (-99, 27.6, -90.8, 33.5) were used to create a bounding box surrounding the Houston metropolitan area, as shown in Figure 7. A total of 6541641 streaming tweets were collected during Harvey. These tweets include a very large number of irrelevant messages. Hence, the data was further filtered using the filtering process outlined in Figure 6. First, duplicates and retweets were removed. Then, the remaining tweets were filtered using a keyword-matching approach. A collection of N crisis terms (keywords) from the CrisisLex lexicon [104] was selected. These terms include disaster-related terms (such as FEMA, Houston, flood, and military), rescue-related terms (such as relief, rescue, evacuate, save, help), and damage-related terms (such as damage, casualties, affected, trapped, search, and shelter). The logical relationship between these terms can be expressed by the following First-Order Logic

(FOL) expression:

$$\forall x\,(\text{hasDisasterKeywords}(x) \vee \text{hasRescueKeywords}(x) \vee \text{hasDamageKeywords}(x))$$

where the universe of discourse is the tweets collected. This expression implies that only tweets containing disaster-related terms, rescue-related terms, or damage-related terms are retained, while all other tweets are discarded.

Location and rescue hashtags are relevant features of the emergency rescue tweets. Without location information, a rescue request is of limited use for first responders. Hashtags such as #SOSHarvey, #SOSHouston, and #Rescue are commonly used to flag rescue requests on social media. Hence, the next step was using *regular expressions* (*regex*) to detect tweets with location patterns and hashtags. The logical relationship between these filters can be expressed by the following First-Order Logic (FOL) expression:

$$\forall x\,(\text{hasLocationFeature}(x) \vee \text{hasRescueHashtag}(x))$$

This expression implies that tweets having certain rescue hashtags, or a location pattern are retained while the remaining tweets are excluded. To identify tweets with location patterns, a relaxed variant of the *regex* expression used for identifying address features (this expression will be introduced in the next section) was employed. The final filtered sample comprised 5,792 tweets, which were manually labeled.

For Hurricanes Ida and Ian, the tweets were collected post-event. Twitter's academic API v2 was used to search for rescue tweets posted during these hurricanes. This API allows users to search for historical tweets using a set of keywords. Several combinations of keywords, including

hurricane names and hashtags (e.g., #Irma, #Ida), flooded zones (e.g., Fort Myers in Florida), and help/rescue keywords and hashtags (e.g., #sos, need help, family stuck), were employed. In total, a sample of 4044 tweets was collected from Hurricane Ida, and another sample of 1,017 tweets was collected from Hurricane Ian. These tweets were also manually labeled.

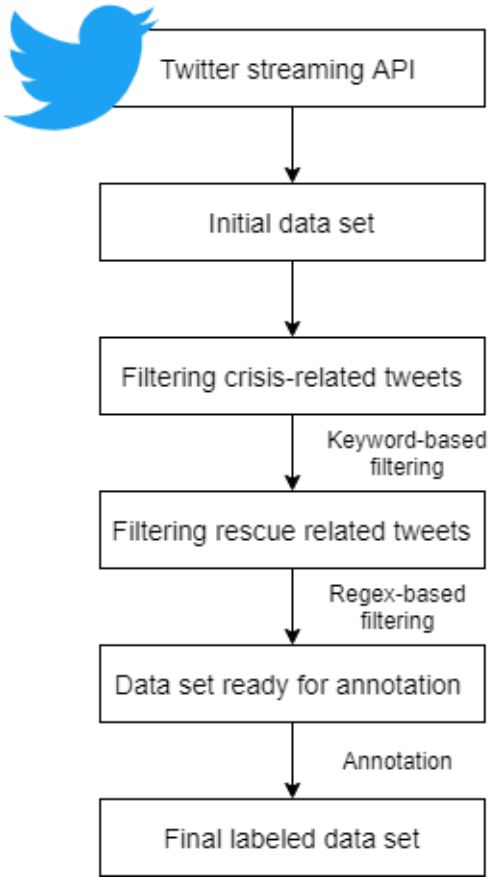

Figure 6. Streaming data filtering workflow for Hurricane Harvey

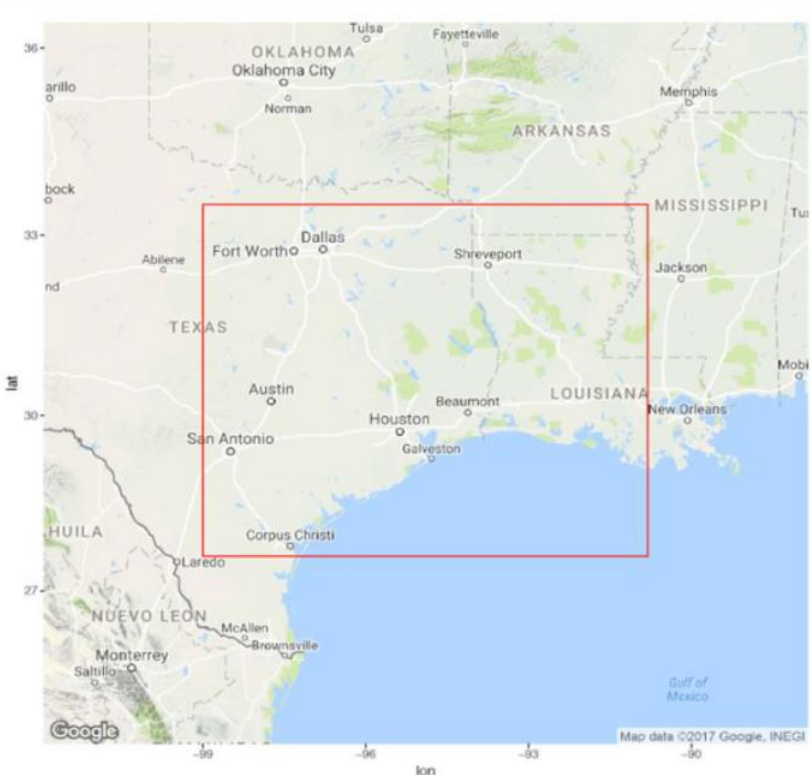

Figure 7. The bounding box used to collect tweets during Hurricane Harvey

3.3.3 Data Annotation

Two categories for labeling the collected tweets were defined: (1) emergency rescue requests (SOS) and (2) non-emergency tweets. The author of this dissertation established two criteria to define the emergency category. **Criterion 1:** For a tweet to be categorized as an emergency rescue tweet, it should include a location reference [144]. Locations can be specified either by using a full US address pattern (e.g., <House Number> <Street Name> <Street Suffix> [<Unit>] <City/Town Name> <State> <Zip Code>) or a location description (e.g., the intersection of street *X* and *Y,* at a gas station in *Z*, etc.). **Criterion 2:** An emergency rescue tweet should include an emergency-related expression or contextual information about the urgent situation. Examples of

emergency-related keywords and hashtags include ‘please help’, ‘needs help’, ‘people stranded’, ‘#SOS’ and ‘#SOSHarvey’ among others. Emergency rescue tweets might include (but are not limited to) additional details about the current situation, such as the person who needs help, the number of stranded people, missing relatives, etc. Figure 8 shows a typical example of an emergency rescue tweet captured during Harvey that meets the above two criteria. It includes a specific U.S. address (‘7815 Pacific Spring Ln’) and a rescue expression (‘please help my friend’). Tweets that did not meet the above criteria were assigned to the non-emergency category.

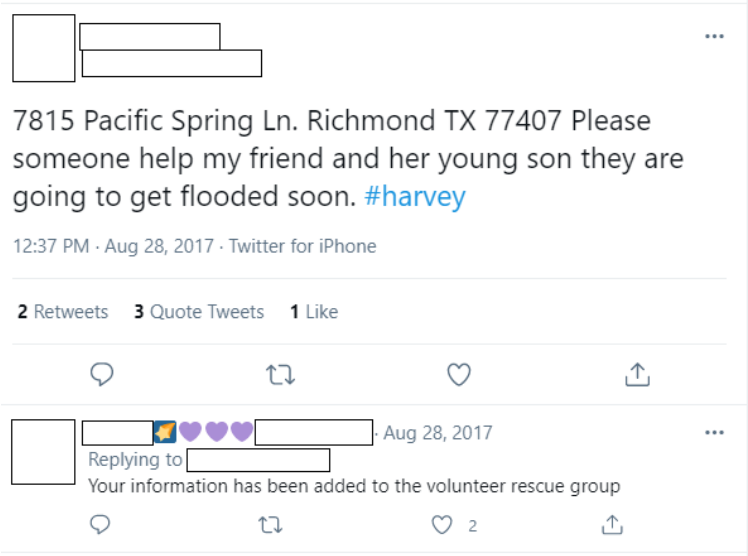

Figure 8. A typical example of an emergency rescue request tweet

After applying the data filtering process, the final dataset from Hurricane Harvey, which included 5,792 tweets, was manually labeled by graduate students as part of their coursework in two graduate-level courses during the Fall 2020 semester. The students were divided into groups of three, with each group assigned a subset of tweets to label. Each tweet was labeled as either an

emergency rescue tweet or a non-rescue tweet. The inter-rater reliability between annotators, measured in percentage agreement, was satisfactory (95.33%). All the annotated tweets were reviewed by the author of this dissertation for further verification. The final dataset collected during Hurricanes Ida and Ian, which included 5061 tweets, was fully labeled by the first author of this dissertation. A subset of 774 tweets was collected, representing approximately 15% of the total data set. This subset was given to an independent annotator to label according to the same annotation guideline described earlier. The inter-rater percentage reliability agreement was calculated on this subset, which was also satisfactory (95.09%).

3.3.4 Proposed Emergency Rescue Requests Features

The proposed features for this study are derived from analyzing the logical structure of the problem and its key characteristics. The proposed features can be categorized into four types: (1) contextual features, (2) location-based features, (3) ask-for-help features, and (4) other features.

Contextual features refer to the information within the rescue message that indicates an urgent situation or gives details about the urgent situation. These features include (1) keywords/expressions describing an emergency (e.g., stranded, trapped, etc.), (2) emergency hashtags (e.g., #SOSHouston, #HelpHouston, etc.), and (3) mentions to individuals in need of help (e.g., father, kids, family, etc.).

As mentioned previously, location is a key feature of actionable rescue tweets. During hurricanes, emergency response agencies repeatedly emphasize the importance of providing a precise address when requesting help. While most people comply, only a small number of people describe their location using street features such as ‘at the intersection of X and

Y', 'at the end of Road X', and so on. Ideally, an emergency rescue tweet includes a full U.S. address. However, some tweets might include a location description instead. Addresses in the U.S. have a clear and simple pattern:

$$< HouseNumber >< StreetName >< StreetSuffix > [< Unit >] \\ < City/TownName >< State >< Zipcode > \quad (1)$$

This pattern can be identified using *regex*. The house number in a US address is usually represented by an integer with up to six digits, with the highest recorded number identified being 107900, as seen in the address '107900 Overseas Hwy, Key Largo, FL 33037'. In the proposed *regex* filter for location identification, the house number was matched as one to six digits. The <street Name> is typically followed by a <Street Suffix> such as 'St' or 'Ave.' The street suffixes listed in the source [138] were used. The *regex* filter for identifying U.S. addresses is defined by the expression in Figure 9:

```
"\\b\\d{1,6}\\s+(#?[A-z]+\\.?(-[A-z]+)?\\s+){1,3}\\b(#@#
|Alley|Allee|Ally|Ally#@#
|Annex|Anex|Annx|Anx|#@#
|Arcade|Arc|#@#
|Avenue|Av|Ave|Aven|Avenu|Avn|Avnue|#@#
…
|Well|Wl|Wells|Wls|#@#
).?\\b"
```

Figure 9. Regex expression for identifying U.S. addresses [51]

The substring ‘d{1,6}’ matches the house number while ‘s+’ matches one or more spaces, including characters for space, tab, vertical tab, newline, carriage return, and form feed. The optional hashtag symbol, which some users use in street names, such as ‘3 friends stuck at 4055 South #Braeswood Boulevard’, is represented by the sub-string ‘#?’ in the *regex* filter. The sub-string ‘(#?[A-z]+.?(-[A-z]+)?s+){1,3}’ represents one to three words consisting of English characters, possibly joined by a ‘-’ or ended with a ‘.’, ‘Alley’, ‘Allee’, ‘Aly’, and others are possible street suffixes, with over 200 possible suffixes listed in the *regex* expression. In the expression shown in Figure 9, only the first few lines and the last line of suffixes are included, with many lines in the middle omitted. It is also possible to see addresses in other forms, for example, ‘1108 Highway 7’ and ‘123 Avenue G’. This address format is matched with the *regex* expression shown in Figure 9.

While the location is a crucial feature for hurricane rescue requests, it alone is not sufficient to identify such requests. Another relevant feature is whether a tweet includes hashtags or keywords about asking for help such as ‘HurricaneRescue’, ‘FloodRescue’, ‘please help’, ‘need to be rescued’, etc. This feature can be captured easily with *regex*. Some tweets possess the above features but are not rescue requests, such as updates on rescue status, offers of help with shelter and food, political tweets, news reports, and commercial tweets.

3.3.5 Proposed Classification Architecture

The proposed architecture consists of two key components: (1) a feature extractor and (2) a Multi-Layer Perceptron (MLP) classifier. The feature extractor transforms input tweets into two distinct feature vectors: (1) low-level textual feature vector and (2) problem-specific high-level

feature vector. These vectors are then combined and used as input to the MLP classifier. Figure 10 illustrates the proposed model's architecture.

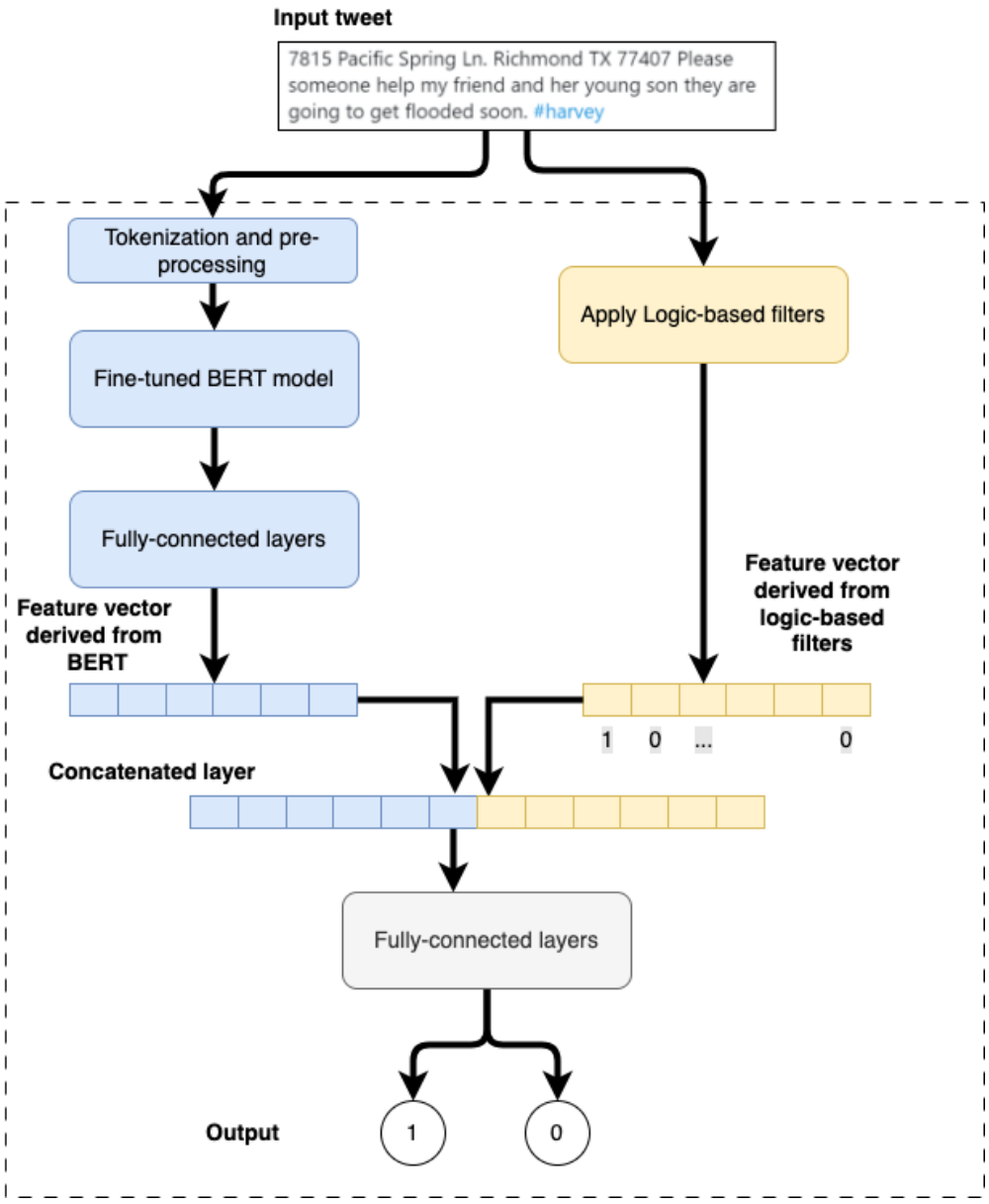

Figure 10. System architecture of the proposed model: For a given tweet, two sets of features were extracted from the tweet, then concatenated and fed into fully connected layers to classify whether the tweet is an emergency request.

3.3.5.1 Feature extraction

The proposed classification architecture combines two distinct types of features: (1) low-level textual features and (2) high-level problem-specific features. This architecture takes in two types of inputs: the textual features produced by a fine-tuned BERT model and high-level problem-specific features generated by *regex*.

High-level feature extraction – A set of *regular expression (regex)* filters was implemented to transform raw input tweets into a *d*-dimensional feature vector. Each *regex* filter identifies specific language patterns or keywords corresponding to a given feature. The elements of the resulting vector are binary, with a value of 1 if the feature is detected and 0 otherwise. Table 2 describes the different filters used in this study. For example, the feature vector for the tweet displayed in Figure 8 is <0,1,1,1,1,0,0,0,0,0>, where each value corresponds to a specific *regex* filter. This tweet includes an emergency hashtag (F1.2), a reference to victims needing rescue (F1.3), a help-seeking expression (F2), and a specific U.S. address (F3.1 and F3.2). The 'other features' (F4, F5, F6, and F7) are not detected in this tweet. The resulting high-level input feature vectors are resized by copying each binary value 30 times.

Low-level feature extraction – Textual features can be extracted using either NLP statistical models, such as TF-IDF and GloVe, or pre-trained transformer models, such as BERT. This study used the pre-trained BERT [35] model for low-level feature extraction. The output of BERT is an abstract, comprehensive encoding of the input text that can be stacked with any classifier to perform the final classification using the feature space created by the transformer encoding. Devlin et al. [35] proposed two strategies to extract document representations from pre-trained BERT: (1) using the last hidden layer, which corresponds to the '[CLS]' token, providing a feature vector

of size 768, or (2) concatenating the outputs of the last 4 hidden layers of the '[CLS]' token into a 3072-dimensional embedding vector. The second approach was used to extract textual feature vectors from BERT.

Table 2. Regex filters used for high-level features extraction

Filter	Type	Description
$F1.1$	Contextual	identifies whether the tweet includes keywords describing the emergency
$F1.2$	Contextual	identifies whether the tweet includes emergency hashtags (e.g., #HarveyFlood)
$F1.3$	Contextual	Identifies whether the tweet includes information about people in need of rescue
$F2$	Help-seeking	Identifies whether the tweet includes rescue/help request
$F3.1$ $F3.2$	Location	Identifies whether a U.S. address pattern exists in the tweet
$F4$	Other	Identify whether the tweet includes political content
$F5$	Other	Identify whether the tweet includes an offer for help
$F6$	Other	Identify whether the tweet includes news reports
$F7$	Other	Identify whether the tweet includes a situational update about a rescue

3.3.5.2 MLP classifier

The proposed integrated classification approach involves a two-stage training process. In the first stage, the BERT model was fine-tuned on the training data. Then, the fine-tuned BERT model weights were frozen, and the entire BERT model (excluding the classification head) was integrated into the proposed architecture. In the second stage, the frozen fine-tuned BERT model was used to produce the textual features. Fully connected layers were added on top of BERT to reduce the dimensionality of its output vector. The logic-based feature vector (obtained by regex filters) was resized by copying each binary value 30 times. These two vectors were concatenated and used as input to the MLP classifier. The MLP classifier consists of a couple of fully connected layers followed by the output layer. The whole architecture was trained end-to-end. The optimal model architecture (i.e., optimal number and sizes of the added fully connected layers, optimal hyperparameters, etc.) is determined in the experiments.

3.3.6 Logic-Based Approach for Identifying Rescue Tweets

This dissertation also introduces a logic-based classification model for identifying emergency rescue tweets using the proposed *regex* filters. These filters are integrated using a logical relationship expressed by the following first-order logic expression 2:

$$\begin{aligned}\forall x(&\text{hasFeatureAddress}(x) \wedge (\text{hasFeatureAskHelp}(x)\\ &\vee \text{hasFeatureDisasterContext}(x))) \wedge \neg(\text{hasFeatureStatusUpdate}(x)\\ &\vee \text{hasFeatureOfferHelp}(x) \vee \text{hasFeatureNewsReport}(x)\\ &\vee \text{hasFeaturePolitical}(x) \vee \text{hasFeatureAds}(x))\end{aligned} \tag{2}$$

where the universe of discourse comprises the collected tweets. The semantics of each predicate are indicated by its name and the features discussed in the previous section. This expression implies that a tweet with a location feature and either contextual or ask-for-help features, and without any of the 'other' features, can be classified as an emergency rescue tweet automatically. By evaluating the truth value of the previously defined logical expression (2), each input tweet can be classified as either a rescue tweet or a non-rescue tweet.

For example, the following tweet was posted during Hurricane Harvey: 'Urgently need #Water- Rescue at 10415 Merry Meadow Ln. Elderly #WaterRescueNeeded #HarveyFlood #houstonflood #hurricaneharvey'. The tweet has a location feature (10415 Merry Meadow Ln.), ask for help key- words (urgently need), a victim in need (Elderly), and a rescue hashtag (#WaterRescueNeeded), but none of the political content, news, rescue status update, and help offer features.

3.3.7 Visualization

After categorizing the tweets into emergency rescue and non-emergency tweets, a visualization module was developed for mapping actionable rescue information using *regex* and GeoPy. First, the location description for each classified rescue tweet was extracted using the *regex* U.S. address pattern [51]. Some of the extracted addresses were incomplete; the GeoPy Python package was used to infer the complete addresses and their latitude/longitude coordinates. The extracted locations were then displayed on a map using the Folium library. Figure 11 shows an example of an interactive map that includes a subset of rescue tweets posted during Hurricane Ian in Florida.

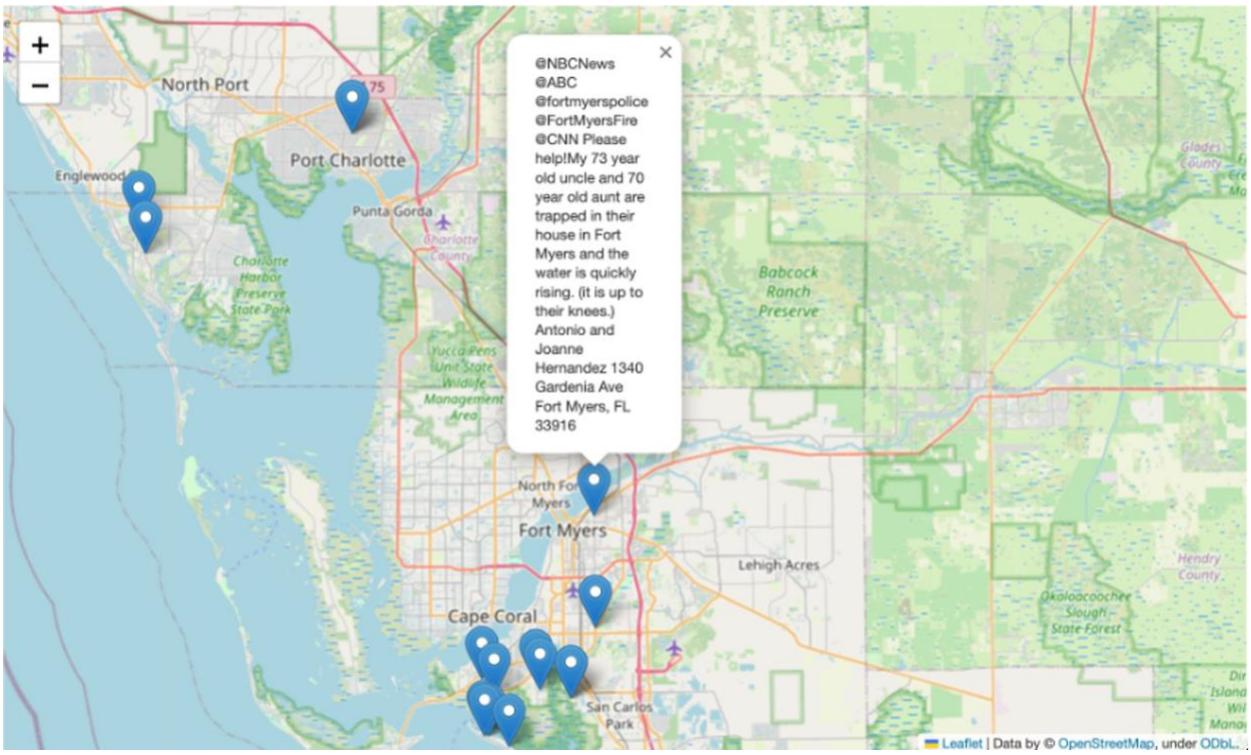

Figure 11. Visualization of a subset of IAN emergency rescue tweets

3.3.8 Competing Methods

The following methods were implemented for comparison: (1) Support vector machine with TFIDF for feature extraction (TFIDF+SVM)[34], (2) VictimFinder architecture [148] that consists of a pretrained BERT model with LSTM head (BERT+LSTM), (3) Convolutional Neural Network with GloVe embedding model (Glove+CNN) [34], and (4) Fine-tuned BERT classifier (BERT+linear).

TFIDF+SVM classifier – Support Vector Machine (SVM) [30] is a robust discriminative supervised learning model that has been successfully applied in a variety of applications, including text classification, face recognition, image classification, and others. The SVM algorithm is trained to

identify an optimal N-dimensional hyperplane (or decision boundary), with N being the number of features, to separate the training data points with a maximum distance (or margin) between classes. This model is very effective in high-dimensional input spaces, such as those encountered in text classification applications, and is also robust against outliers. The SVM architecture is shown in Figure 12. The margin size, specified by the regularization hyperparameter, is the distance between the black lines. The linearity of the hyperplane is determined by the kernel function, such as linear, polynomial, or Gaussian RBF. SVM was employed in the Devaraj et al. [34] study for identifying rescue messages posted during Hurricane Harvey.

The Term Frequency-Inverse Document Frequency (TF-IDF) was used for feature extraction. In this method, a raw tweet is considered a document and converted to a numerical vector with a length N that corresponds to the vocabulary size. TF-IDF is a widely used statistical representation scheme in NLP applications. It takes into consideration the relevance of a word to a specific document and its frequency in the entire corpus. TF-IDF is calculated as the product of two metrics, term frequency (tf) and inverse document frequency (IDF). It is represented by the following equation:

$$TFIDF\ (t, d, D) = tf(t, d) * idf(t, D) \quad (3)$$

where t represents the term, d denotes a document (in this case, a tweet), and D is the collection of documents (corpus). Term frequency can be calculated in various ways, such as raw count, boolean count, or logarithmic scaling. The raw count method (i.e., measuring the frequency of each word in a document) was used in this study. Inverse document frequency (IDF) adjusts the weights assigned to terms based on their frequency across the entire corpus. It is calculated as follows:

$$IDF\,(t, D) = log\,(\frac{N}{|d \in D : t \in d|}) \quad (4)$$

where N is the total number of documents and $|d \in D : t \in d|$ is the frequency of the term t in the corpus.

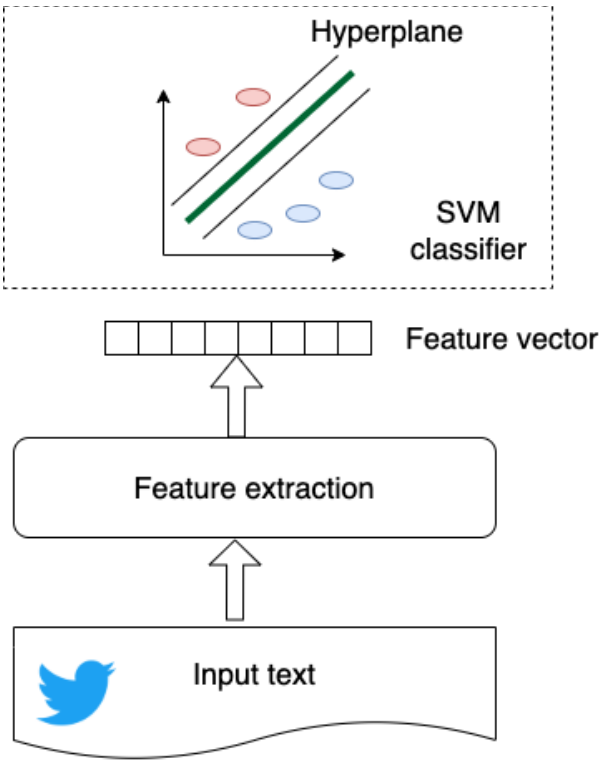

Figure 12. Support Vector Machine (SVM) classifier

GloVe+CNN classifier – Convolutional neural networks (CNNs) are widely used in computer vision and image processing applications and have recently been applied to natural language processing tasks such as text classification. Wide and shallow CNNs have proven to be effective for short text classification, such as Twitter messages. The convolution and pooling operations of CNNs capture spatial information, such as relevant n-grams, in text. This study used the CNN architecture proposed by Kim [64] for classification. The GloVe model [108] was used to extract features from raw tweets. GloVe (Global Vectors for Word Representation) is a log-bilinear, unsupervised model for building distributed word representations. GloVe combines two families of methods: global word co-occurrence matrix factorization and local window methods, such as Skip-grams and Continuous Bag of Words (CBOW). It has been shown that GloVe outperforms other word representation models, including word2vec [93], in various natural language processing tasks, such as word analogy, word similarity, and named entity recognition.

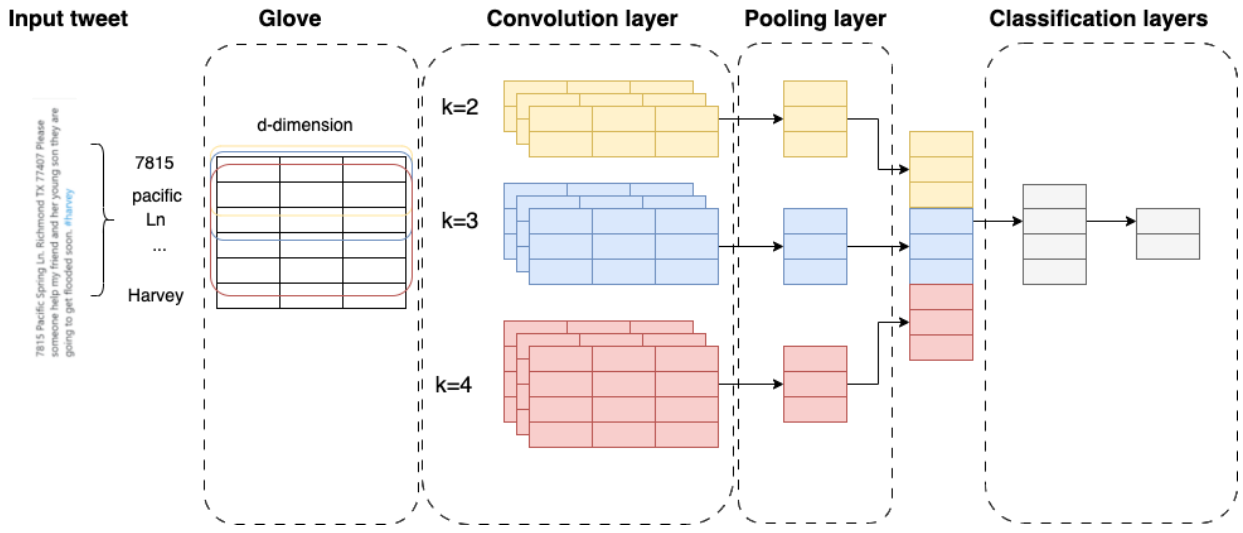

Figure 13. CNN architecture used in this study

CNN has been applied in [96] and [34] to identify emergency rescue messages posted on social media during hurricane events.

BERT+LSTM classifier – VictimFinder is a set of transformer-based models proposed by Zhou et al. [148]. The BERT-LSTM architecture was selected as the backbone for VictimFinder since it achieved the highest performance among the transformer-based models evaluated in [148]. This architecture employs a pre-trained BERT model with a bi-directional LSTM head. This study employed the pre-trained BERT model[3] to extract features from the raw tweets. This model consists of 12 transformer blocks, 768 hidden units per layer, 12 attention heads, and an overall 110 mil- lion parameters, and was pre-trained on a large lower-case English corpus. The 768-dimensional feature vector produced by BERT is fed to a 768-dimensional LSTM layer, followed by the output layer to perform classification.

End-to-end fine-tuned BERT – In addition to the previous models, BERT was fine-tuned end- to-end. BERT (Bidirectional Encoder Representations from Transformer) [35] is a pre-trained transformer network developed by Google. BERT has achieved state-of-the-art performance for several NLP tasks. BERT is typically pre-trained on a large, unlabeled corpus (e.g., English Wikipedia, with up to 2500M words) for masked word prediction and next-sentence prediction tasks. BERT can be fine-tuned for various downstream tasks by adding one or more fully connected layers on top of the core model. The whole model, including the BERT core model and the top- classifier, can be trained end-to-end for a specific downstream task. BERT is released in two general-purpose pre-trained variants:

- $BERT_{base}$: This variant has 12 transformer blocks, 768 hidden size, 12 attention

[3] https://huggingface.co/transformers/v3.3.1/pretrained_models.html

heads, and in total 110M parameters. This architecture is used in this research.

- $BERT_{large}$: This variant has 24 transformer blocks, 1024 hidden size, 16 attention heads, and in total 340M parameters.

The final hidden '[CLS]' was used for the final BERT output. Then, several fully connected layers, followed by an output layer with SoftMax activation, were added on top of the core BERT model. The fine-tuned BERT architecture is illustrated in Figure 14. The optimal number of fully connected layers added on top of the core BERT model and their sizes are determined by Optuna [5] through the experiments. During inference, each input tweet is assigned to the target class with the highest probability.

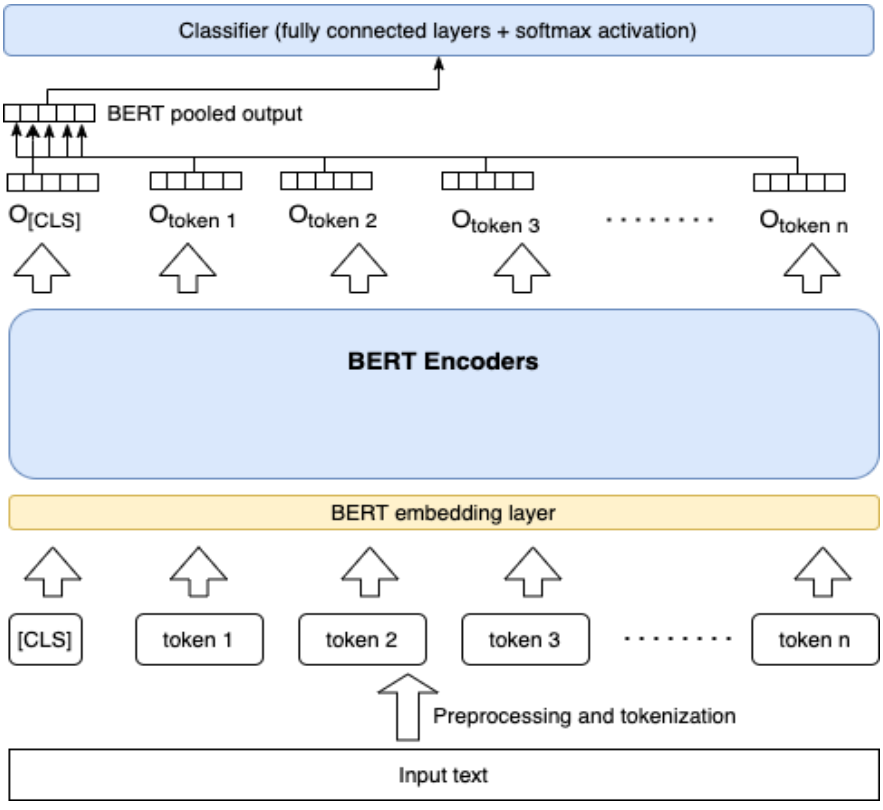

Figure 14. Fine-tuning BERT for classification

3.3.9 Evaluation Metrics

Due to the class imbalance issue, relying on accuracy as a performance metric for quantitative evaluation is inadequate. The following metrics were used to provide a more comprehensive evaluation: precision, recall, F1 score, Area Under the Precision-Recall Curve (AUC-PR), and confusion matrices (CMs).

Precision measures the accuracy of positive predictions by calculating the fraction of correctly identified emergency rescue tweets over the total number of positive predictions. In other words, this metric reports the proportion of positive tweets that were correct.

$$Precision = \frac{TP}{TP + FP}$$

Recall, also referred to as the true positive rate or sensitivity, is the fraction of correctly identified positive tweets (TP) relative to the total number of positives in the dataset. It shows the proportion of actual emergency rescue tweets that were correctly predicted.

$$Recall = \frac{TP}{TP + FN}$$

The F1 score is the weighted average of precision and recall, as shown by the equation below. This metric was selected because it takes into account how data is distributed, making it an effective metric, particularly when the training dataset is imbalanced.

$$F = 2 * \frac{precision * recall}{precision + recall}$$

Since the training data sets are highly imbalanced, the area under the precision-recall curve (AUC-PR) was employed as an additional performance metric for this study. This metric is widely used in information retrieval applications, and it is appropriate for detecting rare events, such as the emergency rescue tweets in this study's case and does not depend on model specificity [33]. AUC-PR ranges from 0 to 1, where a value closer to 1 indicates better model performance. The AUC-PR curve is produced by plotting the precision values against the recall values at various thresholds for binary classification. The AUC- PR score summarizes the overall performance of the model, regardless of the choice of threshold. This makes it a more robust evaluation metric than the F1 score, which assumes a threshold value of 0.5.

Finally, the classification confusion matrices were plotted to compare the performance of the different classifiers. Confusion matrices for binary classification problems are 2X2 tables that indicate the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The basic structure of a confusion matrix is shown in Table 3.

Table 3. Confusion matrix basic structure

	Actual Positive	**Actual Negative**
Predicted Positive	TP	FP
Predicted Negative	FN	TN

3.4 METHODS FOR THE RELIABILITY ASSESSMENT PROBLEM

This dissertation introduces a two-stage rule-based scoring system for assessing the reliability of rescue messages posted on Twitter during hurricane events. Unlike the previous reliability

assessment scoring systems for social media data, the proposed system in this study analyzes the reliability of rescue messages at both an aggregate level (referred to as the claim level) and at the tweet level. The rescue tweets are analyzed based on three reliability dimensions: (1) source-level reliability, (2) post-level reliability, and (3) contextual-level reliability. The reliability assessment methodology proposed in this study involves the following steps:

1. Step 1 – Conduct a literature review to identify the main reliability factors (also referred to as 'reliability indicators' in the remainder of this study) used for analyzing the reliability of social media data. An expert opinion was also used to help select the appropriate reliability indicators.

2. Step 2 – Select a list of reliability indicators to use in the proposed rule-based model. Quantify these indicators using metrics derived from related historical data, recent statistics, and the literature.

3. Step 3 – Develop the reliability assessment model using the selected factors.

4. Step 4 – Evaluate the model using ground truth data–a manually annotated data set of rescue tweets labeled by their reliability.

3.4.1 Data Collection and Annotation

To assess the proposed model's performance, a gold standard (ground truth data) must be built. In previous research studies, the ground truth data sets were manually constructed. Researchers used annotators who were tasked with determining the reliability of given social media messages and judging whether they were reliable or not. To the best of the author's knowledge, no prior data

set has been introduced for evaluating the reliability of social media rescue information during natural disasters. Hence, this study built an annotated data set from a subset of rescue tweets collected in the first part. The following annotation procedure was defined. Initially, the collected rescue tweets from the first problem were organized into several ‘claims’ based on the locations indicated in the tweets’ text. Tweets matching the same location (e.g., a U.S. address or a precise location description) were grouped into a single claim. For each claim, the Twitter search bar was used to find all available rescue tweets sharing the same ‘claim’. The outcome of these two steps is a set of rescue claims, each of which corresponds to a specific location. Each claim is shared by one or more tweets, either from the same or different users. An example of a rescue claim is presented in Figure 15. The tweets in this example describe a rescue situation at ‘8502 Elm St’, thereby forming a unique rescue claim at this location.

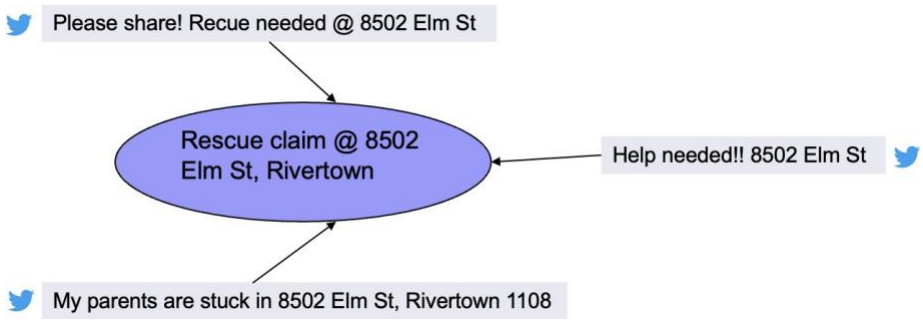

Figure 15. An example of a rescue ‘claim’

Afterward, each ‘claim’ was manually verified and categorized using the following annotation taxonomy. Rescue messages were categorized into either: (1) high-reliability claims or (2) low-

reliability claims. The FEMA Historical Geospatial Damage Assessment Database[4] was used for annotation. This database serves as a repository for geospatial damage assessments from past national disaster events, conducted using either high-resolution imagery or geospatial modeling [41]. An example of a rescue claim that is located in an impacted FEMA zone is illustrated in Figure 16. In this figure, the address specified in the claim is located in an area where numerous structures, such as houses and buildings, were marked as damaged (affected, minor dam- age, major damage, and destroyed) structures. For each rescue claim, the number of 'damaged' structures within a 0.3-mile radius of the claim's address was analyzed. Claims surrounded by a large number of 'damaged' structures were marked as 'high-reliability'. Claims that were not surrounded by damaged structures (or surrounded by only a very few of them) were marked as 'low-reliability'.

As previously noted, it is very difficult to assess the veracity of rescue messages posted on social media in the aftermath of a hurricane. The information posted on social media can be easily verified in some cases; for instance, when someone posts a piece of misinformation, rumor, or fake news, external sources, such as fact-checking websites, might be used to assess the veracity of the information and debunk it. However, it is more difficult to certify a call for help or a rescue request posted on social media unless one is present at the moment of the call [150]. Therefore, the reliability of the collected rescue messages was approximated by comparing the information to an external (official) source. This study assumes that a rescue message is likely to be true if it originates from an area impacted by the disaster, such as regions flooded during hurricanes.

[4] https://experience.arcgis.com/experience/c1b507827e72401aace5a6d277fad93b/page/Page-1/?views=Visual-Damage-Assessments

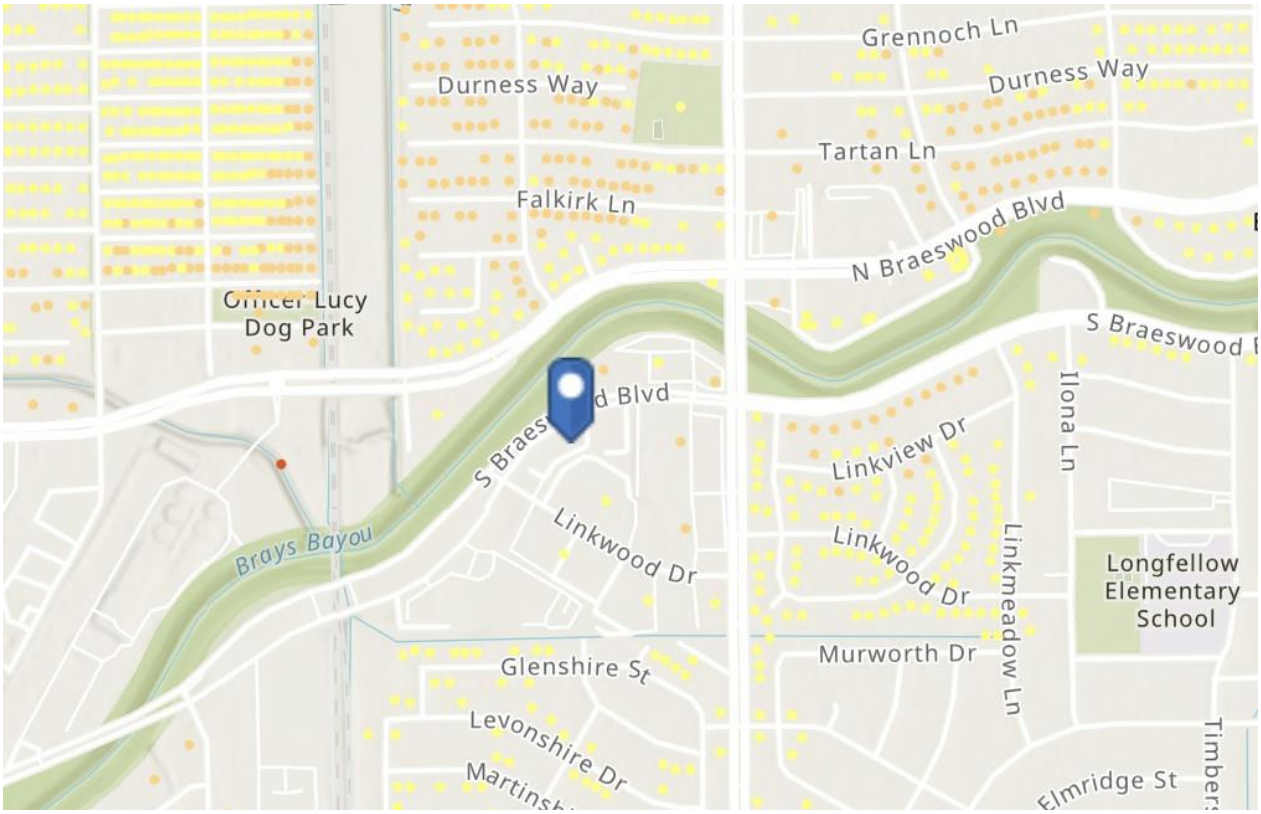

Figure 16. A rescue claim situated in a FEMA impacted zone

Validation with an external source is a form of triangulation, where the information is corroborated with other sources. Disaster response professionals [144] highlighted that they often rely on triangulation—i.e., getting the same information from multiple sources— to determine the reliability of a social media report. This study investigates whether the proposed reliability indicators can accurately predict the rescue claims coming from the damaged areas as a proxy for reliability.

3.4.2 Proposed Reliability Scoring System

The second contribution of this research is a two-stage framework for prioritizing ‘actionable’

rescue information on Twitter based on their reliability. After filtering emergency rescue tweets (previous part), the framework processes the extracted information in two stages: (1) reliability assessment of individual tweets (stage 1) and (2) intelligent aggregation and claim-level confidence scores calculation (stage 2). The framework structure is depicted in Figure 17.

The framework employs a scoring mechanism to assign a reliability score for each incoming tweet. As new tweets are posted, the scoring is updated in real-time. The tweet-level reliability score is calculated based on two assessment dimensions: (1) source-level assessment and (2) post-level assessment. Intuitively, a social media message with a high-reliability score indicates that the user who posted the message is trustworthy and that the content is of high quality (e.g., attached image/video, high content interaction, etc.). Once the tweet-level reliability scores are calculated, an intelligent aggregation of the tweets is performed. Tweets that refer to the same information (indicated by a given U.S. address and/or precise location) are grouped to form an actionable rescue 'claim' defined as follows:

> *An actionable rescue claim k, denoted as C_k, encompasses a call for help or a rescue message that may trigger a potential 'action'. A claim is associated with a specific location, such as a U.S. address, and may be supported by either a single tweet or multiple tweets.*

For each claim, a confidence score, denoted as conf(C_k), is derived from the reliability scores of its related tweets. At this stage, contextual indicators, such as location and corroboration, are integrated. Intuitively, a claim that is posted at a location where a high number of rescue-related messages are posted and is supported by many tweets (contextual indicators) is likely to be true, thereby getting a higher confidence score.

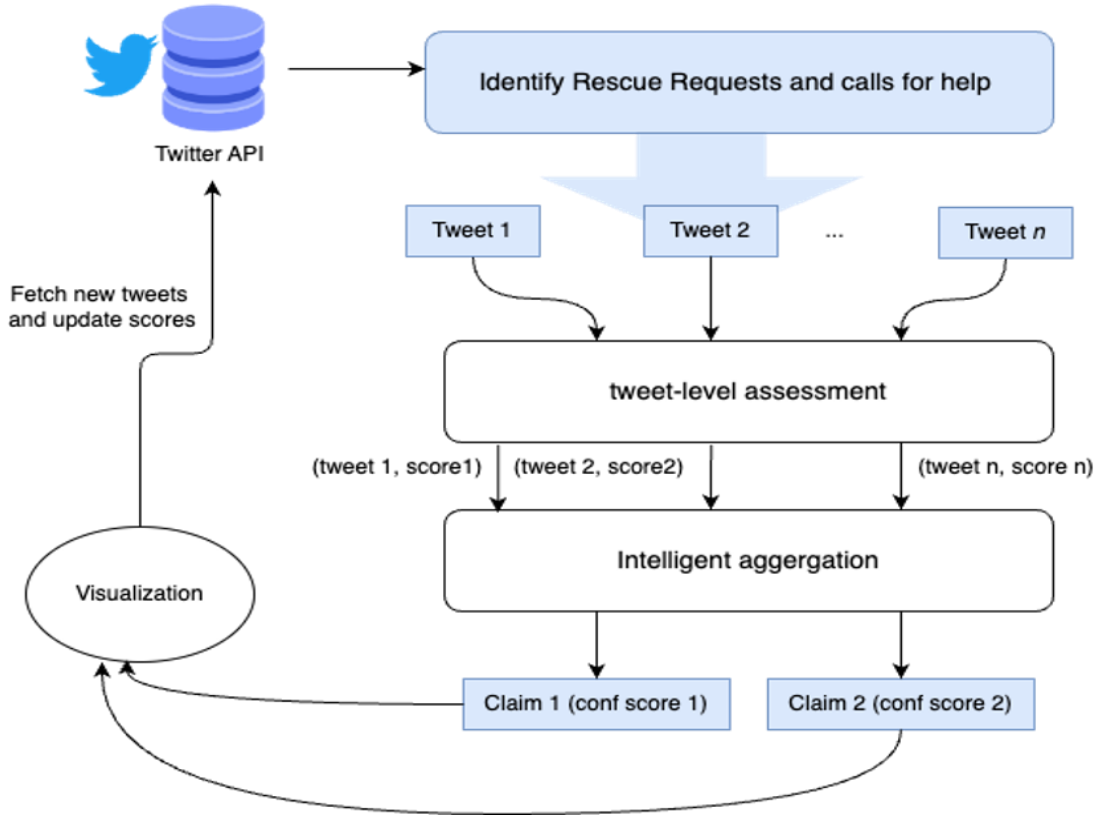

Figure 17. reliability assessment framework

All the components of the proposed scoring system are linked together in an algorithmic form to analyze the reliability of a given rescue claim. The novel features of the proposed reliability assessment framework are outlined as follows. Unlike most of the previous studies that have formulated the problem as a binary classification problem, this study proposes a reliability scoring system where reliability is modeled by a numerical confidence score between 0 and 1. Most of the research on assessing the reliability of social media data has adopted the machine learning approach to categorize social media messages and potentially identify suspicious social media messages. The systems most closely related to the one proposed in this dissertation are the reliability scoring systems introduced by Alrubaian et al. [11] and Assery et al. [16]. The

framework developed by Alrubaian et al. [11] includes three modules designed to assess the reliability of the content, user reputation, and user expertise, operating together algorithmically to calculate a final reliability score for each tweet. Assery et al. [16] proposed a reliability assessment model that analyzes disaster-related tweets based on post and user-related dimensions. The reliability score provided by this model is calculated on a 10-point scale. Each of these frameworks assesses the reliability of individual tweets. The proposed reliability scoring framework in this study differs from these models by introducing a 'claim-level' reliability that aggregates tweets' level reliability scores. Furthermore, the previous reliability assessment systems focused on content-level and user-level reliability features, the proposed reliability assessment model in this study integrates contextual features. However, it is important to note that the model's performance was evaluated on a binary scale, like previous research, given the current challenges in obtaining ground truth data labeled by scores. Because having this type of label requires extensive involvement from humanitarian and disaster response experts in the annotation process, this evaluation approach is deferred to future research.

3.4.2.1 Definition of reliability

Numerous studies have attempted to define 'reliability;' however, it has been noted that no clear definition has been proposed so far [109]. Researchers have defined the reliability of social media information through various criteria, including relevance, trustworthiness, and expertise [109]. The Merriam-Webster dictionary defines reliability as 'the quality or power of inspiring belief' or 'capacity for belief.' The Oxford Dictionary defines it as 'the quality of being believed'. In this study, the latter definition was adopted, which has been frequently utilized in prior studies [11][109]. Based on this definition, reliability was modeled as a confidence score ranging from

0 to 1, reflecting the perceived level of belief in a piece of information based on several indicators about its content, source, and context. The confidence scores assigned to a rescue claim is defined as follows.

> *The confidence score (conf(C_k)) of a claim C_k is a numerical score assigned to a rescue claim k, ranging from 0 to 1, that reflects the degree to which the information conveyed by the claim can be believed.*

Similarly, the confidence scores assigned to a rescue tweet is defined as follows.

> *The confidence score (conf(T_i)) of a tweet T_i is a numerical score assigned to the tweet, ranging from 0 to 1, that reflects the degree of belief in the information it conveys.*

3.4.2.2 Reliability indicators selection

A set of reliability indicators was selected for the proposed reliability assessment model. These indicators were derived from the literature and an informal discussion with an expert. In the literature, reliability has been analyzed from different levels [109] [58] [2] [23]: (1) medium level, (2) topic level, (2) source level, (3) post level, and (4) contextual level.

While social media magnified the issue of verifying the 'accuracy' of recorded information (such as information posted on books, newspapers, and websites), the challenge predates the social media era. Fallis [38] identified four key areas to consider when verifying the accuracy of recorded information. The first suggested area to consider is the authority of the information source. Source authority can be determined through several factors, such as expertise (or reputation) and the history of the author in providing reliable information. The second area is the independent corroboration [150]; it involves validating claims through complementary information from other

sources (e.g., several reliable independent websites supporting a piece of medical information). As indicated by Fallis [38], it is much more likely that one individual will 'deceive or be deceived' than that several individuals will 'deceive or be deceived' in the same way. The other two areas, plausibility and presentation, assess the context surrounding the information and how it is presented, respectively. In [123], journalists and editors from various news media outlets (e.g., BBC, CNN, NBC) published a step-by-step process for verifying online-generated content. To verify a piece of information on social media, journalists and humanitarian professionals should check the following elements: (1) provenance (is the information original?), (2) source (is the source that uploaded the content trustworthy?), (3) date (when was the content created?), and (4) location (where was the content created?). The journalistic approach also encompasses three levels of verification: (1) verifying the source of the information, (2) verifying the information itself (such as information provenance), and (3) verifying the context in which the information was posted, particularly the date and location of the information. This study conducted a literature review of the related problems to select a set of reliability indicators.

More than 60 studies were reviewed in total. The reviewed studies covered various social media reliability assessment tasks, including fake news, misinformation, and rumor detection. The search scope included not only the disaster relief domain but also general news, politics, health, and other areas. During the review process, studies proposing black-box machine learning and deep learning models relying solely on raw textual features were excluded. The list of papers initially reviewed was narrowed down to 36 studies that used specific user-related, content-related, or contextual features in their reliability assessment methods. The identified features were grouped into different categories (reliability constructs). The final list of the selected reliability indicators is

illustrated in Table 4. The selected indicators include source reputation, source authenticity, source expertise, source location, content engagement, direct evidence, proximity, and corroboration. More calculation details of the reliability indicators assessment are provided in Appendix C.

Source reputation – Measuring source reputation is an important aspect of reliability. Similar to [11] and [125], reputation in this study was approximated by the popularity of the social media account, leveraging features, such as the relationships between following, followers, and friends. Legitimate accounts usually follow known users who follow them back. The number of followers is often almost equal to the number of followees [125]. An exceptionally high number of followers, ranging from tens of thousands to millions, indicates that the account is a celebrity, thereby implying a high reputation. Conversely, accounts with a very low number of followers but a large number of followees are indicative of suspicious behavior. They tend to connect to as many users as they can to spread their messages through the network. In this dissertation, the reputation of a social media account posting a rescue message is calculated by the average of two popularity metrics.

> *The reputation of a user j, denoted by $Reputation_j$, is defined as a numerical score ranging from* 0 *to* 1. *This score is computed as the average of two Twitter Follower-Followee ratios, namely TFF^1 and TFF^2.*

Table 4. Selected reliability indicators

Indicator	Dimension	Description
Authenticity	Source	This indicator analyzes the legitimacy of a social media profile.
Reputation	Source	This indicator analyzes the perception of a social media account based on its behavior. Messages shared by suspicious user accounts tend to be unreliable.
Location	Source	This indicator analyzes the user's known location. Accounts that are geo-tagged in the vicinity of the crisis zone area are more likely to be direct eyewitnesses, thereby having a higher reliability level
Expertise	Source	This indicator analyzes the user's level of knowledge and skills. Users with direct expertise, skills, or roles related to the disaster are generally trustworthy
Engagement	Content	This indicator analyzes the popularity of the message and the level of engagement that it generates.
Direct evidence	Content	This indicator analyzes the presence of direct evidence (an image or a video) attached to the social media post
Cross-checking	Context	This indicator analyzes the extent to which the shared information is supported by external sources
Proximity	Context	This indicator analyzes the spatial context of the shared information, that is, how many urgent messages are posted from the proximity of the post's location.
Corroboration	Context	This indicator analyzes external sources of data to confirm the posted call for help. The FEMA flooding risk map was used as an external source

The first Follower-Followee ratio TFF^1 is given in Eq. 5. This metric calculates a score ranging from 0 to 1. TFF^1 values that are close to 1 indicate that the user is popular, whereas reputation scores close to 0 are indicative of potentially suspicious users.

$$TFF^1 = \frac{nb.followers}{nb.follower + nb.followees} \quad (5)$$

To calculate the second Follower-Followee ratio TFF^2, this study starts from $Assery_{TFF}$ metric (Eq. 6).

$$Assery_{TFF} = \frac{nb.followers}{nb.followees} \quad (6)$$

$Assery_{TFF}$ is used by Assery et al. [16] for measuring profiles' popularity on Twitter. They categorized Twitter profiles into 5 categories based on this ratio (as shown in Table 5): (1) Spammer category, (2) Suspicious user category, (3) Normal user, (4) Micro influencer, (5) influencer.

Table 5. Classification of users based on follower/following ratio

[16] Score	Category
less than 0.5	Spammer
between 0.5 and 1	Suspicious
between 1 and 2	Normal
between 2 and 10	Micro Influencer
higher than 10	Influencer

Based on a given user's profile type, the second Twitter Follower-Followee ratio TFF^2 is calculated as follows.

$$TFF^2 = \begin{cases} 0.25 & \text{if user } j \text{ is spammer or suspicious} \\ 0.75 & \text{if user } j \text{ is normal} \\ 1 & \text{if user } j \text{ is micro influencer or influencer} \end{cases} \quad (7)$$

Source authenticity – Twitter's 'verified' status was used as an indicator of source authenticity. Twitter marks 'verified' accounts with blue checkmarks. Previously, the 'verified' status is assigned once a given account meets several eligibility requirements, such as authenticity (e.g., ID verification, official website, etc), notability (e.g., associated with a predominantly recognized individual or brand), and user activity. However, since April 2023, these criteria have been updated. To receive the blue badge, a profile now simply needs to be complete (e.g., displaying a name and a photo), has a confirmed phone number, be active, and show no deceptive signs (e.g., stable activity without spam indicators) [25]. Despite the current eligibility criteria for verification being less rigorous, the 'verified' status of an account is still considered a strong indicator of the social media user's authenticity. This research investigated the distribution of verified accounts among bots and legitimate accounts in two publicly available bot detection data sets (see Appendix A). As shown in Figure 18, among 9,481 bot profiles in both data sets, only 2 of them were verified. However, among 10,741 legitimate profiles, 453 were verified. This analysis suggests that almost all verified accounts are authentic.

> *The authenticity of a user j, approximated by their verification status and denoted by* $Verified_j$*, is defined as a binary value. This score is assigned a value of* 1 *for verified users and* 0 *for non-verified users.*

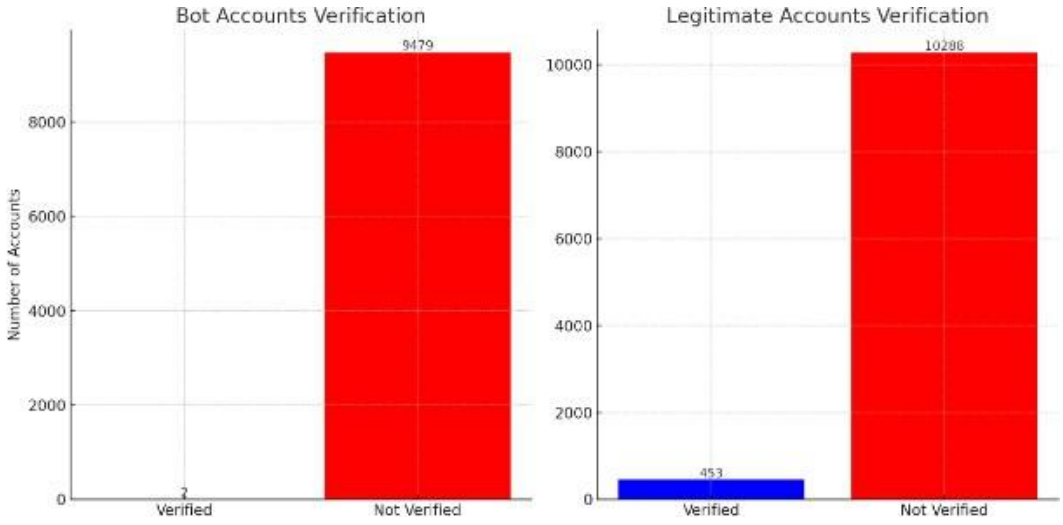

Figure 18. Number of verified profiles among bots and legitimate accounts

Source Expertise – The expertise is a domain-dependent indicator. Expertise refers to the level of knowledge and skills that a social media user possesses in a particular field or area. A profile on Twitter is categorized as an 'expert' if it belongs to the following categories: (1) officials, (2) journalists, (3) emergency response (affiliated) volunteers, and (4) meteorologists. Profiles not fitting these criteria are classified under the 'public' category.

> *The expertise of a user j, denoted by* $Expertise_j$*, is defined as a binary value. A value of* 1 *is assigned for expert users and* 0 *for non-expert users.*

Source Location – Twitter's geo-tagging feature allows users to tag their locations. Although only about 14% of social media users share their location information via geo-tagging, this feature has been used as a reliability predictor in several studies. The distribution of geotagged users among legitimate and bot accounts in the publicly available bot detection datasets (see Appendix

A) was analyzed. As shown in Figure 19, of the 9,481 bot profiles captured in both datasets, only1,073 were geo-tagged (11.31%). In contrast, of the 10,741 legitimate profiles, 5,035 were geo- tagged (46.87%). There is a significantly higher number of geo-tagged users among legitimate accounts compared to bot accounts. Further details on how the geotag value is calculated can be found in Appendix C.

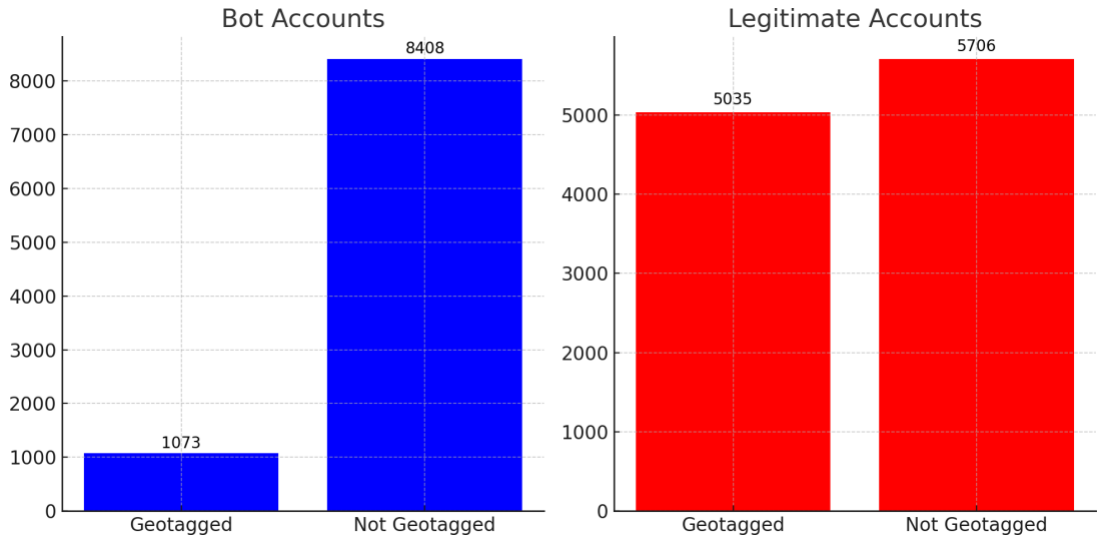

Figure 19. Number of geo-tagged profiles among bots and legitimate accounts

The location of a user j, denoted by $GeoTag_j$, is indicated by their use of the Twitter geo-tagging feature. $GeoTag_j$ is assigned a value of 0 *if the user is not geotagged within the disaster area, and* 0.84 *otherwise.*

Content Engagement – The content-related indicator measures the level of interaction a social media message receives over a specified period. Key engagement features frequently cited in the literature include the number of retweets, likes, and replies to a post. The content engagement

score is defined as follows.

> *The content engagement of a tweet i, denoted by E_i, measures the level of interaction received by a post. It is calculated by the average of two tweet engagement metrics: TE^1 and TE^2.*

The first tweet engagement metric TE_i^1 is an adaptation of the Twitter engagement ratio [5]which is a crucial metric in Twitter analytics. This metric assesses a profile's ability to engage and reach its audience. It is calculated by dividing the engagement metrics (retweets, likes, replies, and quotes) by the number of posts made by the user. This metric has been employed as a reliability indicator in various studies, such as [16] and [117]. The Twitter Analytics engagement is given by Eq. 8.

$$AnalyticsEngagement_i = \frac{nb.retweets + nb.replies}{nb.followers}$$

Then, depending on the value of $AnalyticsEngagement_i$, an engagement category was assigned in the following way:

- Low engagement ($AnalyticsEngagement_{C1}$): retweets and replies to follower, ratio lower than 0.081.
- Mild engagement ($AnalyticsEngagement_{C2}$): retweets and replies to followers, ratio between lower 0.081 and 0.7483
- Medium engagement ($AnalyticsEngagement_{C3}$): retweets and replies to followers, ratio between lower 0.7483 and 1.333.
- High engagement ($AnalyticsEngagement_{C4}$): retweets and replies to follower, ratio

[5] https://scrunch.com/blog/what-is-a-good-engagement-rate-on-twitter

higher than 1.333.

The first tweet engagement metric TE^1 is calculated as follows.

$$TE^1 = \begin{cases} 0.7035 & \text{if the tweet belongs to the low engagement category} \\ 0.6883 & \text{if the tweet belongs to the mild engagement category} \\ 0.7464 & \text{if the tweet belongs to the medium engagement category} \\ 0.5540 & \text{if the tweet belongs to the high engagement category} \end{cases} \quad (9)$$

These values are derived from analyzing the data set published by Assery et al. [16]. The steps used for calculating these values are described in Appendix C.

Furthermore, the dataset proposed by Assery et al. [16] was analyzed to investigate the relationship between the popularity of tweets (measured by the number of retweets) and their assigned reliability level. The retweet counts were divided into four groups based on the second quantile (Q2), third quantile (Q3), and the 90th quantile, as shown in Table 6. Four categories were created for retweet count groups:

- Low popularity (rt_{C1}): 0 to 3 retweets
- Medium popularity (rt_{C2}): 4 to 17 retweets
- High popularity (rt_{C3}): 18 to 95 retweets
- Very high popularity (rt_{C4}): more than 95 retweets

Table 6. Distribution of number of retweets ([16] data set)

Quantile (ϑ^{th})	nb.retweets
0.25	0
0.5	3
0.75	17
0.90	95

Based on the popularity class assigned to a tweet i, the second tweet engagement value TE_i^2 given to the tweet is calculated as follows. The steps employed to calculate these values are also explained in Appendix C.

$$TE_i^2 = \begin{cases} 0.6269 & \text{if the tweet belongs to the low popularity} \\ 0.7925 & \text{if the tweet belongs to the medium popularity} \\ 0.7215 & \text{if the tweet belongs to the high popularity} \\ 0.8166 & \text{if the tweet belongs to the very high popularity} \end{cases} \quad (10)$$

Consider the following scenario: a tweet i has 5 retweets, 2 replies, and 10 followers. The $AnalyticsEngagement_i$ is calculated as 0.7, derived from $(\frac{5+2}{10})$. Consequently, the tweet is categorized under the mild engagement category. Therefore, the first engagement value TE^1 is equal to 0.6883. With 5 retweets, this tweet is classified under the medium popularity category. Consequently, TE^2 is determined to be 0.7925. The final content engagement value for this tweet, E_i, is calculated as 0.7404, representing the average of the previous two metrics.

Direct evidence – This content-related indicator focuses on analyzing the videos and images attached to the social media message. Typically, the presence of direct evidence, i.e., real-time videos and images depicting the incident or situation, would enhance decision-makers confidence in the reliability of the shared information. Nevertheless, the authenticity and trustworthiness of these attachments are crucial. An attached image or video that is fake or reused from past events can substantially undermine the message's reliability. The authenticity of the attached images and videos in the annotated data set by this study was manually verified by the author of this dissertation.

> *The direct evidence attached to a tweet i, denoted as* $Attachement_i$*, is a binary value that takes the value of* 1 *if an authentic image or a video is attached to the tweet's body and* 0 *otherwise.*

Proximity – proximity is a context-related indicator that focuses on assessing the geographical vicinity of the message's location and analyzes the volume of similar urgent messages posted within that area. When a message originates from a hotspot area—i.e., a location characterized by a high frequency of emergency messages—this factor would significantly enhance confidence in its reliability. An example of how to calculate proximity is shown in Figure 20.

> *The proximity of a rescue claim* C_k*, denoted as* $Proximity_k$*, is a numerical score that ranges from 0 to 1. This score is calculated by a sigmoid function that takes in the number of emergency messages within a circular geographical area with a radius r surrounding the claim's location. The output score increases to reach 1 as a higher number of rescue messages are posted on Twitter within this area.*

For each actionable claim *k*, a circular geographical area with a radius r surrounding its location.

Figure 20. Proximity example – how many emergency rescue tweets are posted within the vicinity of the rescue claim in red? The vicinity is represented by a circle with a radius *r*

was defined. The number of tweets within this area was calculated. The proximity score P_i of a tweet *i* is given by the following equation (Eq. 11):

$$P_i = \frac{1}{1 + e^{-n+3}} \tag{11}$$

Where *n* is the number of messages posted within the specified area. As shown in Figure 21, as a higher number of messages are posted in the vicinity of a given call for help claim, this score increases to reach 1 at a certain point.

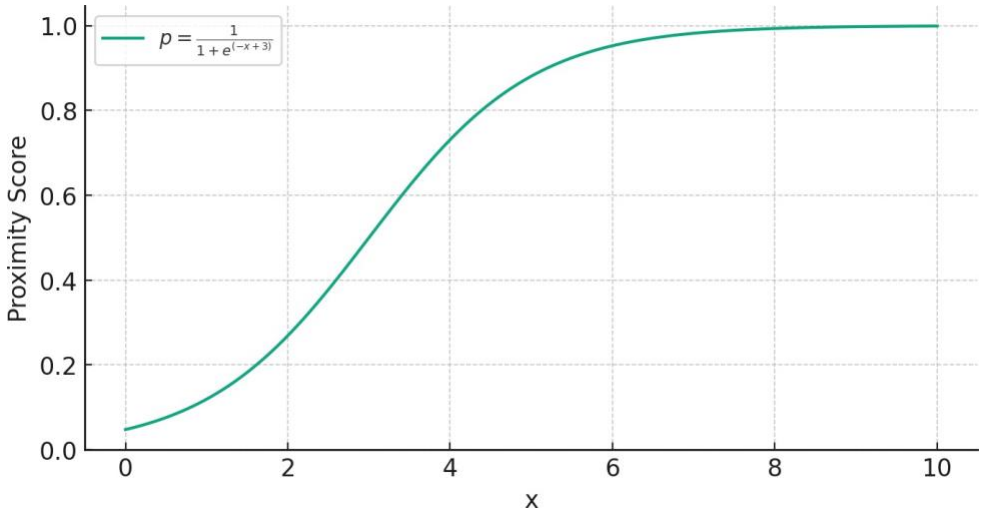

Figure 21. Proximity score calculation – number of tweets vs proximity score

Corroboration – Corroboration (with external sources) is a context-related indicator that analyzes external sources to find evidence that supports or refutes a rescue claim. In this study, the FEMA flooding risk map[6] was used as an external source for corroborating a given rescue claim. FEMA defines several flood risk zones, including (1) high-risk zones (zones AE), and (2) low-risk zone (Zone X). The risk zone feature is derived from the FEMA flood zone designations. FEMA flood maps (Flood Insurance Rate Maps FIRMs) are efficient tools provided by FEMA for under- standing flood risk across a geographical area. These maps are used to provide information about the local flood risk and determine the necessity for flood insurance. FEMA flood zones can be categorized into high, moderate, and low flood zones. Each zone reflects the severity of flooding in a particular area. High-risk areas are usually labeled with letters A or V (e.g., AE, AO, AH, A1-A30, VE). These areas have a 1% annual chance of flooding and a 26% chance of flooding over a 30-year mortgage. If a property owner has a

[6] https://hazards-fema.maps.arcgis.com/apps/webappviewer/index.html?id=8b0adb51996444d4879338b5529aa9cd

federally backed mortgage, flood insurance is mandatory. Moderate to low-risk areas are usually labeled by B, C, or X. Although these areas are outside the 1 annual flood risk zone, they may still be at risk from flooding. There is no federal requirement for flood insurance in these zones, but it is recommended. An example of a FEMA flood zone is depicted in Figure 22, where the blue area represents a designated FEMA 'AE' zone (high-risk), and the brown shaded area represents a designated FEMA 'X' zone (low-risk).

Figure 22. FEMA flood risk zones (example)

For each rescue claim k, the designated zone (e.g., zone AE, zone X, etc) was identified. Subsequently, this information is converted into a risk penalty value v. This value is set to $-v$ if the claim is located in a low-risk flood zone and to $+v$ if it is located in a high-risk flood zone. This risk value is then used to adjust the confidence score assigned to the claim (as will be explained in the model description section).

> *Corroboration for a rescue claim C_k, denoted as $Corroboration_k$, is a numerical score that represents whether an external source confirms or refutes the claim. $Corroboration_k$ can take either a positive risk value v or a negative risk value $-v$.*

3.4.2.3 Model description

As described earlier, the proposed reliability scoring model has two stages. In the first stage, a confidence score ($con\,f\,(T_i)$) is calculated for each tweet. In the second stage, an intelligent aggregation of tweets' confidence scores is performed to produce claim-level confidence scores ($con\,f\,(C_k)$).

Tweet-level assessment (stage 1) consists of two components as illustrated in Figure 23: (1) source-level assessment and (2) content-level assessment. This assessment is performed periodically, and the scores assigned to each tweet/user are updated over time. The final tweet-level score is a weighted average of source and content, as shown in Equation 12. A weight w_u was assigned for the source-related score, and another w_t was assigned for the content-related score.

$$Con\,f\,(T_i) = w_u * U^i_{cred} + w_t * P^i_{cred} \tag{12}$$

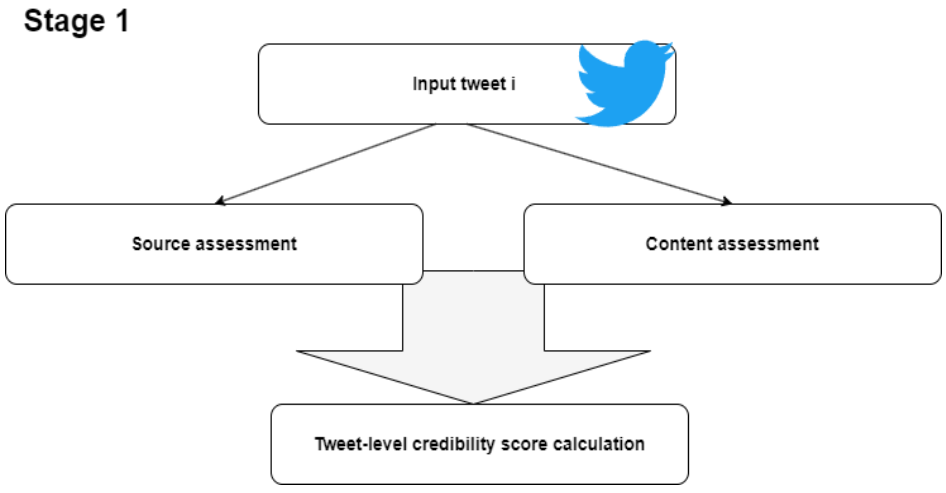

Figure 23. reliability score calculation for an input tweet *i*

The source (user) reliability score, denoted by U_{cred}^{j} is calculated using algorithm 1. Initially, the model checks if the Twitter account is 'verified'. Verified users are automatically assigned a score of 1. The model then evaluates the user's expertise, with accounts identified as 'expert' also receiving a source reliability score of 1. For public and non-verified profiles, the model calculates the user's reputation score $Reputation_j$. Additionally, the model checks if the user is geo-tagged; if so, the reputation score $Reputation_j$ is adjusted to reflect the enhanced reliability from being geo-tagged within the disaster area. If not geo-tagged, the source reliability score is set equal to the calculated reputation score $Reputation_j$.

Algorithm 1 Calculate User reliability

```
function CALCULUSERCRED(user_j)
    if Expertise_j = 1 OR Verified_j = 1 then
        U^j_cred ← 1
    else
        Calculate reputation score Reputation_j
        if GeoTag_j = 0 then
            U^j_cred ← Reputation_j
        else
            U^j_cred ← Reputation_j * (1 + GeoTag_j)
    return U^j_cred
```

The content reliability score of a tweet *i*, denoted by P_{cred}^{i} , is calculated using Algorithm 2. Initially, the model determines whether a video or image is attached to the tweet. If an image or video is uploaded, the content reliability score is automatically set to 1. If there is no media attached, the system then proceeds to calculate the engagement score of the post, which corresponds to the final content reliability score. The second stage involves an intelligent aggregation of the tweet- level reliability scores to create a set of 'actionable' rescue claims, as shown in Figure 24.

Algorithm 2 Calculate Content reliability

```
function CALCULCONTENTCRED(tweet_i)
    if Attachement_i = 1 then
        P^i_cred ← 1
    else
        Calculate Engagement score E_i
        P^i_cred ← E_i
    return P^i_cred
```

The calculation of claims' confidence scores follows the basic principle of truth discovery algorithms [140]. A claim's confidence score is calculated by following these steps.

Step 1: In the first step, for each claim, the model calculates the sum of tweets-level reliability scores as shown in Equation 13.

$$\sigma(claim_k) = \sum_{i \in W(t_i)} conf(T_i) \tag{13}$$

Where $W(t_i)$ represents a set of tweets forming the rescue claim C_k.

Step 2: In the second step, the model adjusts the claim's confidence score $\sigma(C_k)$ by adding contextual indicators to the sum. The adjusted reliability score is calculated as follows (equation 14).

$$\sigma^*(C_k) = \sigma(C_k) + I_{context} \tag{14}$$

$I_{context}$ includes the proximity score ($Proximity_k$) and the corroboration score ($Corroboration_k$) assigned to the claim k.

Step 3: In the last step, the model computes the final confidence score to be assigned to claim *k*, denoted by ($con\ f\ (C_k)$). A sigmoid function is applied, as outlined in Equation 15, to normalize this score, resulting in a final confidence value between 0 and 1.

$$conf(C_k) = \frac{1}{1 + e^{-\gamma.\sigma^*(C_k)}} \tag{15}$$

Consider the following scenario: a claim *k* is confirmed by three tweets. The confidence scores assigned to tweets 1, 2, and 3 are 0.5, 0.6, and 0.5, respectively. Summing their confidence scores, $\sigma\ (C_k)$ was equal to 1.6. Subsequently, the values of contextual indicators are integrated. Suppose this claim is located in a FEMA 'AE' zone (low-risk zone). A claim from a low-risk flooding zone is generally considered less credible. This evidence should decrease the confidence assigned to the claim. Assuming a penalty value of 0.5 is selected, the confidence assigned to the claim is penalized by 0.5, resulting in a new $\sigma\ (C_k)$ of 1.1. Additionally, assuming there are 5 rescue messages posted within a predefined radius of 0.3 miles. The $Proximity_k$ value is then set to 0.9674. The $\sigma\ (C_k)$ is adjusted again by 0.9674, resulting in 2.0674. The last step involves normalizing the score using the sigmoid function defined in step 3 of the algorithm, leading to a final $conf\ (C_k)$ of 0.6957. Note that the penalty score significantly reduces the confidence score assigned to the claim, as it is assumed that it is uncommon for a claim to be

posted from an area not susceptible to flooding. However, given the number of people sharing the tweet and other contextual factors (proximity), the claim still receives a relatively high confidence score, making it reliable.

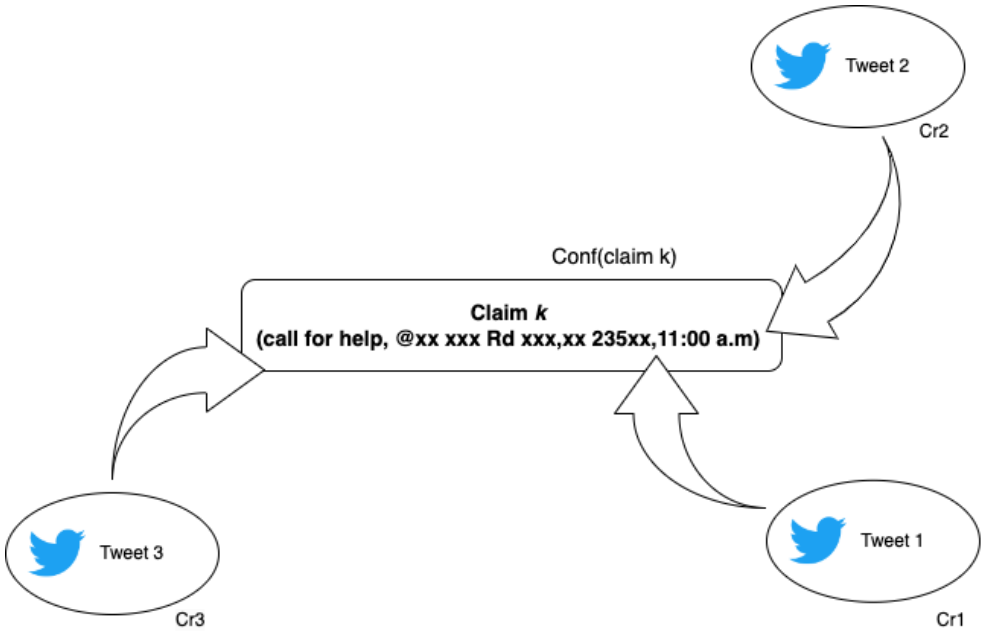

Figure 24. Claim reliability scoring

3.4.3 Competing Methods

To the best of the author's knowledge, no model has been directly designed to address the reliability assessment of rescue messages on social media during hurricane events. Therefore, the model proposed by Assary et al. [16] was used as a competing method. Assary et al. [16] introduced an unsupervised learning model to assess the reliability of disaster-related Twitter data. This model was evaluated using historical tweets from two hurricane events (Hurricanes Michael and Florence) and demonstrated promising performance when compared with a set of

commonly used supervised machine learning models. The model employs a 10-point scoring system to evaluate the reliability of tweets based on both user-based and content-based features. Initially, user-based features are extracted, followed by content-based features. Depending on the final score, the model categorizes a tweet as ‘reliable’ if the obtained score ranges from 5 to 10 and as ‘not credible’ if the score is below 5. The features utilized by this model include:

- Verified user account: if the user’s account is verified by Twitter, ten points are assigned to the tweet, and no further analysis is required.
- Trusted username: The tweet’s score is increased by one point if the username or description contains trusted information sources. In the study’s implementation, trusted sources included journalists, media-related profiles, and weather experts.
- Slang and swear words in the profile description: if the username or description of the profile has slang and swear words, one point is deducted from the tweet’s score. Otherwise, one point is added to the score.
- Follower-following ratio: the popularity of the user is measured via this ratio (Eq. 16). Users are categorized into different popularity classes based on the obtained value. Spammers and suspicious users have a ratio below 1. Normal users have a ratio between 1 and 2. Micro- influencers and influencers typically have a ratio higher than 2. One point is deducted from the scores of the tweets posted by Spammers and suspicious users, one point is added to the scores of the tweets posted by normal users, and three points are added to the scores of the tweets posted by micro-influencers and influencers’ users.

$$Follower/Following = \frac{nb.followers}{nb.following} \tag{16}$$

- URL validity: the proposed model checks the validity and trustworthiness of the URLs in the tweet. Two points are added to the scores of the tweets with trustworthy and valid URLs, while one point is deducted from the scores of the tweets with non-valid URLs. In this context, a valid URL is a URL coming from a trustworthy domain.

- Slang and swear words in the tweet: the model verifies whether the tweet contains slang and swear words. If it does, one point is deducted from the tweet's score. Conversely, one point is added to the scores of the tweets with no slang or swear words.

- Question marks and exclamation: one point is deducted from the tweets, which include exclamation and question marks. One point is added to the tweets with no questions and exclamation marks.

- Tweet engagement ratio: finally, the model calculates a value, denoted by the engagement ratio (Eq. 17). Tweets are categorized into several engagement types. If the value of the ratio is below 0.02%, the tweet has a low engagement rate. If the value of the ratio is between 0.02% and 0.09%, the tweet is considered to have a mild engagement rate. If the value is higher than 0.09%, the tweet has a high engagement rate. One point is deducted from the scores of the tweets with low engagement rates. One point and three points are added to the scores of the tweets with mild and high engagement rates, respectively.

$$Engagement = \frac{nb.likes + nb.retweets}{nb.posts} \quad (17)$$

In addition to the Assery model [16], the proposed reliability assessment model was compared to a selected set of supervised machine learning models. Supervised machine learning is a common approach for addressing social media reliability assessment problems. Given the small size of the annotated data, traditional supervised learning models were selected for comparison, including ensemble machine learning models (e.g., Random Forest and AdaBoost), probabilistic machine learning models (Naive Bayes), and discriminative machine learning models (Decision Trees and Logistic Regression), for comparison instead of deep learning models. With more data becoming available, deep learning models will be explored in future studies.

The Decision Tree (DT) algorithm is a non-parametric supervised learning method suitable for both classification and regression tasks. Decision Trees learn simple decision rules inferred from the data features and make predictions for new instances based on these rules. The structure of a Decision Tree consists of a tree-like model, with the top node serving as the root. This structure is recursively split into decision nodes from the root to the leaf (terminal) nodes. The DT structure is illustrated in Figure 25 with a simple example. In this example, an instance with three features (X1=12, X2=-1, X3=10) is classified based on a set of learned rules. At the first level, the DT algorithm checks if the feature X1 is less than 10. At the second level, it evaluates the value of feature X2. If X2 is less than 0, it then evaluates feature X3. The final classification

decision is made at each leaf node in the tree. Unlike other black-box machine learning models, the Decision Tree algorithm provides a more interpretable classification outcome based on its rule-based approach.

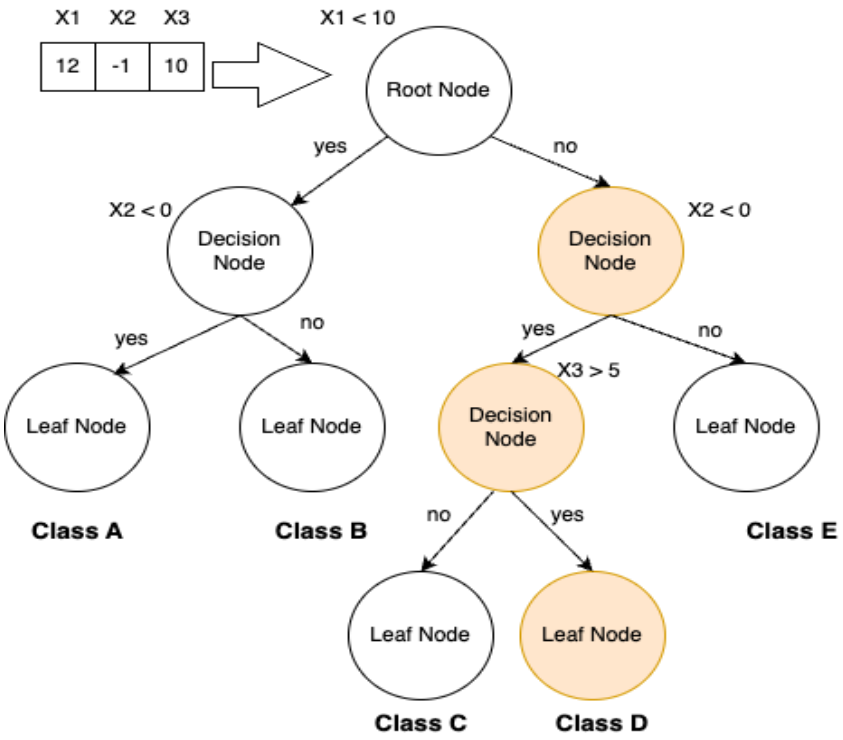

Figure 25. Decision Tree (DT) structure

Random Forest (RF) is another widely used ensemble learning classifier. The fundamental components of a Random Forest classifier are decision trees. An RF model con- sists of multiple decision trees operating together, each employing a subset of randomly selected attributes from the training data to make decisions. New instances are classified through a majority voting mechanism. By integrating multiple weak classifiers, specifically decision trees, the model substantially reduces classification variance, thereby enhancing the reliability of the ensemble

model. For optimal performance, several parameters must be set, primarily the number of trees and the number of randomly selected features. The RF structure is illustrated in Figure 26.

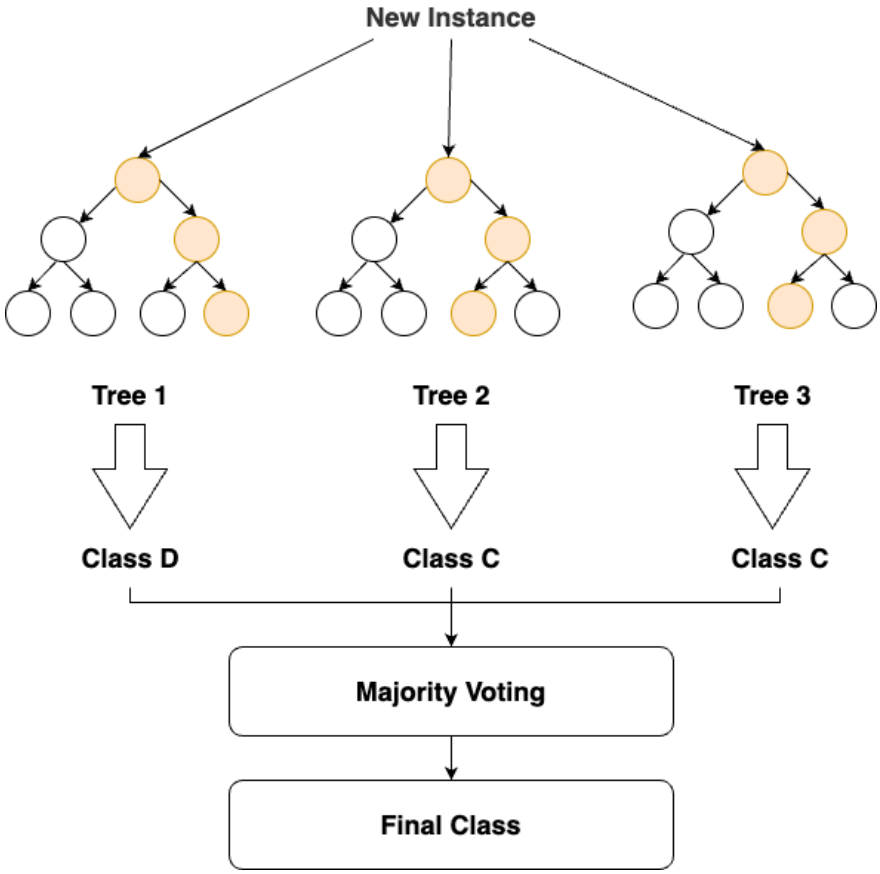

Figure 26. Random Forest (RF) structure (adapted from [65])

AdaBoost (ADA) is a widely used ensemble learning classifier that combines multiple weak classifiers to produce a strong classifier, thereby achieving more accurate classification results. The weak (base) classifiers are trained on various subsets of the training data. The AdaBoost classifier belongs to the category of boosting algorithms, which involves creating a sequence of weak classifiers. Each classifier in the sequence aims to correct the misclassification errors of its predecessors. Consequently, AdaBoost assigns a weight to each training sample,

allocating higher weights to misclassified samples. This mechanism helps to reduce the variance of the classification results. The AdaBoost architecture is illustrated in Figure 27.

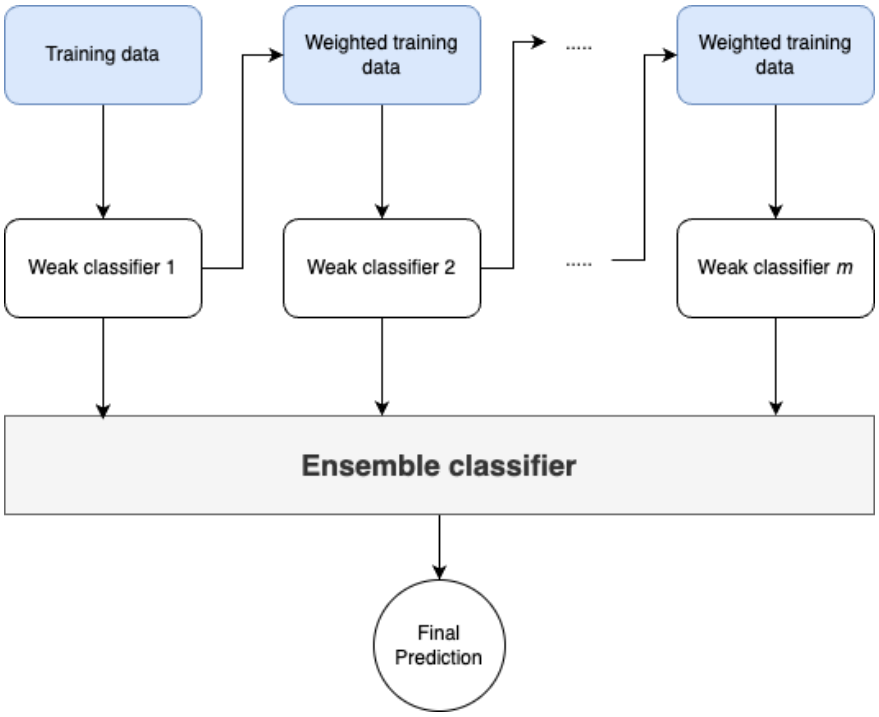

Figure 27. AdaBoost architecture (adapted from [115])

Naïve Bayes (NB) is a probabilistic classification model that applies Bayes' theorem (Eq. 18). The term 'Naïve' refers to the assumption of independence among the input features in the classification process—i.e., it is assumed that each feature contributes independently to the outcome. The classifier calculates the probability of each feature occurring within each class and identifies the most likely class. This model is very effective for classification tasks.

$$P(A|B) = \frac{P(A|B)xP(A)}{P(B)} \tag{18}$$

Logistic regression (LR) is a supervised learning algorithm that is particularly useful for binary classification problems. It operated by predicting the probability of an instance belonging to one class or another. LR uses a logistic function—traditionally the sigmoid function—that takes as input a set of independent variables (features) and outputs a value between 0 and 1. LR is easy to implement and interpret compared to other ML models.

3.4.4 Evaluation Metrics

The proposed reliability assessment model was evaluated using various metrics. Let TP_i, TN_i, FP_i, and FN_i be the number of true positives, true negatives, false positives (type-I error), and false negatives (type-II error) for a class i, respectively. The positive class refers to the 'high-reliability' category, while the negative class refers to the 'low-reliability' category. The equations for the evaluation metrics are given below.

The accuracy metric measures the proportion of correctly classified claims out of the total number of claims in the data set. This metric is useful to determine the degree of correctness of a model. ACC_i represents the accuracy of a class i, whereas $Macro - Acc$ represents the overall accuracy among the two classes.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{19}$$

The precision metric measures the accuracy of positive predictions for a given class (e.g., claims predicted to be 'high-reliability') by taking the fraction of correctly labeled claims over the total number of positive predictions. The precision metric measures the accuracy of positive predictions for a given class (e.g., claims predicted to be 'high-reliability') by taking the fraction of correctly labeled claims over the total number of positive predictions.

In other words, this metric reports the proportion of positive claims that were correct. P_1 refers to the precision related to the 'high-reliability' class, whereas P_0 refers to the precision related to the 'low-reliability' class.

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{20}$$

The recall metric, also referred to as true positive rate or sensitivity, is the fraction of correctly identified positive claims (TP) for a given class to the total number of positive instances in the data set. R_1 refers to the recall related to the 'high-reliability' class, whereas R_0 refers to the recall related to the 'low-reliability' class. The recall of the high-reliability class (R_1) shows the proportion of actual high-reliability claims that were correctly predicted.

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{21}$$

The F-score is the weighted average of the precision and recall metrics, as shown by the equation below. The F1 score provides a balanced evaluation of the classifier's performance for a given class *i*. It captures the tradeoff between precision and recall.

$$F1_i = 2 * \frac{precision_i * recall_i}{precision_i + recall_i} \tag{22}$$

The false alarm rate (also called false positive rate) measures the proportion of low-reliability claims that are incorrectly classified as high-reliability claims.

$$FPR_i = \frac{FP_i}{FP_i + TN_i} \tag{23}$$

The true negative rate, also called specificity, measures the pro- portion of actual true negatives –in this case, low-reliability claims– that are correctly identified by the model.

$$TNR_i = \frac{TN_i}{TN_i + FP_i} \tag{24}$$

CHAPTER 4

RESULTS AND ANALYSIS

This chapter presents the experimental results of this study, which is divided into two sections. The first section details the experimental results for the social media rescue requests identification problem, while the second section discusses the results of the credibility assessment problem.

4.1 RESULTS FOR THE RESCUE MESSAGES IDENTIFICATION PROBLEM

In this section, the following methods were evaluated: (1) the logic-based classification approach and (2) the integrated classification model.

4.1.1 Experimental Setup

4.1.1.1 Data sets

The experiments were conducted using two datasets: (1) the Harvey dataset and (2) the Ian/Ida dataset. The Harvey dataset includes the annotated tweets collected during Hurricane Harvey, while the Ian/Ida dataset includes annotated tweets collected from Hurricanes Ida and Ian, merged into a single dataset. The class distribution in each dataset is reported in Table 7. The total number of tweets manually analyzed in this research was 10853. Of these tweets, 407 rescue tweets were identified.

Table 7. Class distribution in the training data sets

Category	Harvey data set	Ian/Ida data set
Emergency rescue class	272	225
Non-rescue class	5520	4936
Total	5792	5061

4.1.1.2 Pre-processing step

For each input tweet, stop words and punctuation were removed, and all tweets were normalized to lowercase. Numerals were retained in the preprocessing step because they are a crucial part of specifying U.S. address specifications, which is a relevant feature for the problem under consideration.

4.1.1.3 Handling class imbalance

The collected datasets were highly imbalanced. The emergency rescue class (positive class) represents only a small fraction of the annotated tweets. To address this problem, the scikit-learn model selection package was used across all experiments to create balanced training and test sets by maintaining the same proportion of emergency rescue tweets in both. For deep learning models, a weight is assigned to each class in the loss calculation. These weights are based on the data distribution so that each class has a weight proportional to its number of samples, helping to ensure a fair representation of each class during training.

Table 8. Grid search for SVM+TFIDF hyperparameters

Hyperparameter	Configuration space	Description
Kernel	"linear","rbf","poly",	Kernel function
C	1,5,10,15,20, 50,100,150	Regularization parameter
γ	0.001,0.1,1,5,10	Kernel coefficient
Degree	2,3	Polynomial kernel degree

4.1.1.4 Hyperparameter optimization

The Sklearn grid search package[7] was used to select the best-performing Support Vector Machine (SVM) model. Grid search performs a comprehensive search over a defined hyperparameter space—a subset of manually specified hyperparameter values—to identify the optimal model. This optimal model is the combination of hyperparameter values that yields the highest performance according to a given scoring metric. The hyperparameter space used for the SVM experiments is defined in Table 8.

Optuna [5] was used to perform model selection and search for hyperparameters for the proposed integrated classification model and all competing methods except the VictimFinder model, for which the hyperparameters reported in Zhou et al. [148] were used. Optuna is a hyperparameter optimization tool that automates the hyperparameter search in machine learning and includes several modules such as study, storage, trial, sampler, and pruner. The study module controls the tests (also called trials in Optuna terminology) and finds the optimal combination of hyperparameters over a predefined search space. A trial in Optuna is a single evaluation of an

[7] https://scikit-learn.org/stable/

objective function with a set of hyperparameters. In addition to hyperparameter values such as learning rate, batch size, and the number of epochs, the search space in Optuna might include parameters of the evaluated architectures, such as the number of layers and their sizes. The sampler module in Optuna is responsible for generating a set of hyperparameters to try in each trial during the optimization process. Optuna employs several sampling algorithms, such as 'random sampler' which generates random configurations, and 'TPE Sampler' which uses a Bayesian optimization algorithm. The pruner module cuts off trials that are unlikely to yield good performance to reduce the total number of trials. By default, Optuna uses Bayesian optimization to balance exploration (visiting new areas in the search space) and exploitation (focusing on the areas expected to include the optimal configuration). The default Bayesian search algorithm was used in this dissertation. The Optuna search spaces for BERT+Linear and GloVe+CNN competing methods are reported in tables 9 and 10, respectively. The maximum number of trials for all experiments was set to 350. For each combination of hyperparameters, a 10-fold cross-validation was conducted on each dataset. The average AUC-PR was calculated over the 10 folds, and the best combination for each classification method was kept.

4.1.1.5 Experiments

To evaluate the performance of the proposed model, the following experiments are conducted. In the first experiment, the performance of the logic-based approach was evaluated. In the absence of labeled data, this logic-based approach for detecting rescue messages might be helpful in identifying emergency rescue tweets. This experiment aims to evaluate this approach's effectiveness in identifying such tweets. The second experiment compared the proposed integrated classification model with the competing methods. Given the small size of the

manually labeled data sets, 10-fold cross- validation was used. It is expected that the proposed integrated classifier will achieve a superior performance.

Table 9. Optuna search space for GloVe+CNN hyperparameters

Hyperparameter	Configuration space	Description
Kernel	2,3,4,5,6,7,8,9,10,12,16,14,20,24,32	Kernel size
#Filters	16,32,64,128,256,512,1024	Number of kernels
Pool	2,3,4,6,8	Pooling size
Lr	0.1,0.01,0.001,0.005	Learning rate
B	16,32,64,128	Batch size
#hidden units	64,128,256,512,1024	Size of the hidden layer

Table 10. Optuna search space for BERT+Linear hyperparameters

Hyperparameter	Configuration space	Description
Lr	0.00005,0.00003,0.00002	Learning rate
Epochs	4,10,20	Number of training epochs
B	8,16,32,64	Batch size
#layers	1,2,3	Number of layers on top of the BERT backbone
#Hidden units	1024,512,246,32	Sizes of the layer on top of the BERT backbone

4.1.2 Evaluation of the Logic-Based Approach

For the Harvey dataset, the obtained F1 score was 0.8149, recall (sensitivity) was 0.8419, precision was 0.8149, true negative rate (specificity) was 0.9889, and the Matthews correlation coefficient (MCC) was 0.8059. For the Ian/Ida dataset, the obtained F1 score was 0.7897, recall (sensitivity) was 0.6844, precision was 0.9333, true negative rate (specificity) was 0.9977, and

the Matthews correlation coefficient (MCC) was 0.7918. The confusion matrices on Harvey and Ian/Ida data set are shown in figures 28 and 29, respectively. Table 11 summarizes the obtained results for the positive class (i.e., emergency rescue request class) on both data sets. Overall, these results indicate a promising outcome by this model. Tweets with false negative errors obtained by both models are typically those that did not provide complete U.S. addresses, such as "@KPRC2 there are stranded families at Creech Elementary on Mason Rd. You have boats nearby. Please send them!". The low recall obtained by the logic-based model on the Ian/Ida data set is due to the higher number of tweets with incomplete and fuzzy location addresses in this data set compared to the Harvey data set. The regex address expression employed by this model typically detects complete U.S. address patterns that include full house numbers, street names, and suffixes. The outcome of this experiment shows the potential benefits of this approach in supporting the hurricane emergency response in the absence of labeled data sets.

Table 11. Logic-based approach results for the positive class

	F1	Recall	Precision	TNR	MCC
Harvey data set	0.8149	0.8419	0.7896	0.9889	0.8059
Ian/Ida data set	0.8295	0.7244	0.97023	0.9989	0.832

4.1.3 Evaluation of the Proposed Deep Learning Architecture

This section reports the 10-fold cross-validation results on the annotated data sets. The predictions generated by each classifier were further analyzed to understand the behavior of the different models.

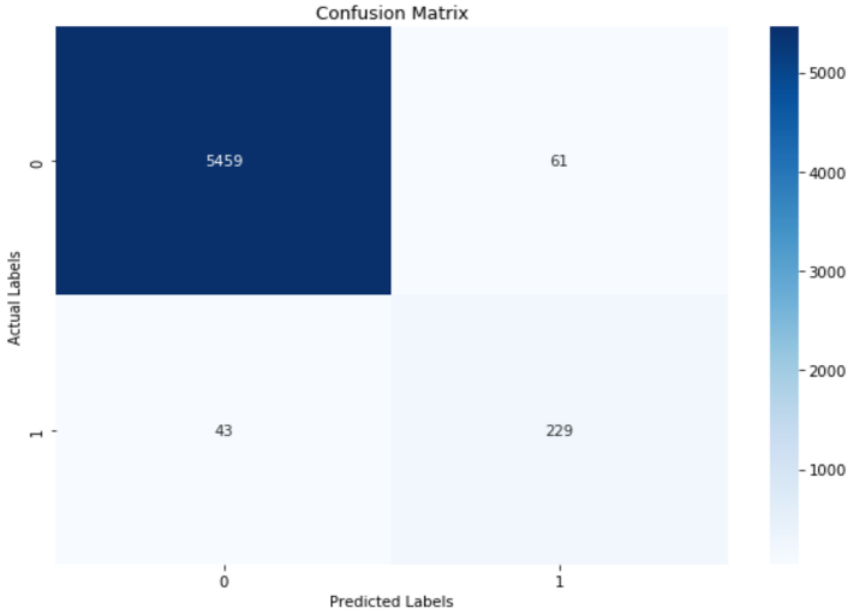

Figure 28. Confusion matrix for logic-based approach on Harvey data set

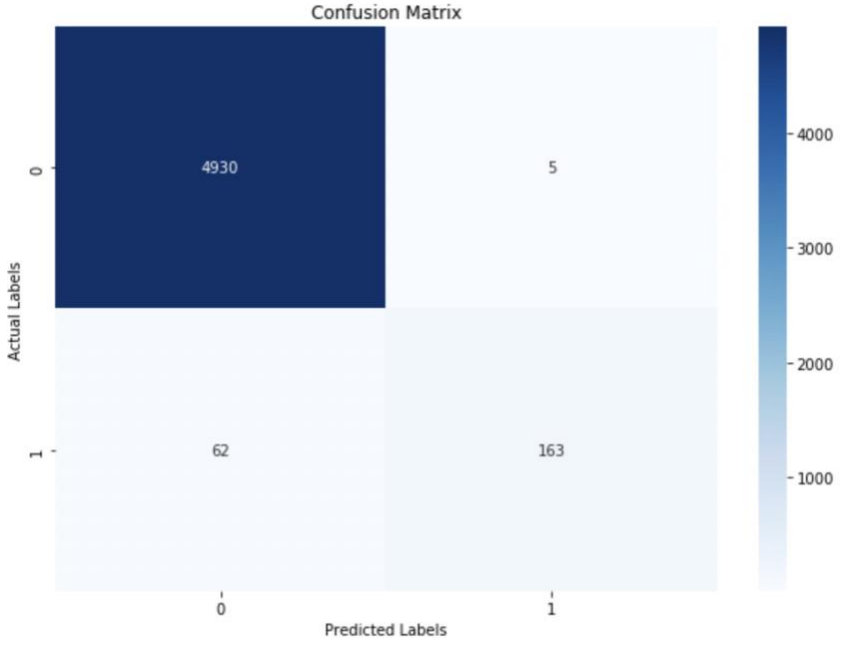

Figure 29. Confusion matrix for logic-based approach on Ida/Ian data set

4.1.3.1 Cross-validation on Harvey dataset

The 10-fold cross-validation results on the Harvey data set are shown in Table 12. This table reports the means (± standard deviations) of the AUC-PR, F1, Recall, and Precision metrics. A paired *t*-test was conducted to compare the AUC-PR scores obtained by the BERT+linear model and those by the proposed integrated model. Table 13 reports the AUC-PR results for each testing fold in the 10-fold cross-validation for the Harvey dataset. The prediction results from the 10 testing folds were concatenated and plotted the AUC-PR curves for the different models, as shown in Figure 30.

Table 12. 10-fold cross-validation results on Harvey data set

Classifier	AUC-PR	F1-score	Recall	Precision
TFIDF+SVM [34]	0.9256 ± 0.03	0.8646 ± 0.03	0.8201 ± 0.05	0.9165 ± 0.04
BERT+LSTM [148]	0.9114 ± 0.08	0.8511 ± 0.03	0.8531 ± 0.04	0.8817 ± 0.007
GloVe+CNN [34]	0.9218 ± 0.03	0.8470 ± 0.04	0.7755 ± 0.08	0.9414 ± 0.03
BERT+Linear	0.9251 ± 0.05	0.8834 ± 0.02	0.9191 ± 0.04	0.8527 ± 0.04
Integrated classifier	0.9621 ± 0.02	0.9093 ± 0.02	0.8898 ± 0.05	0.9329 ± 0.03

The proposed integrated classifier outperformed all the competing models in AUC-PR (0.9621) and F1 score (0.9093). It is worth noting that the most indicative metric is the AUC-PR, which provides an overall performance measure of the model, corresponding to the area under the PR curve shown in Figure 30, where it can be clearly seen that the proposed model has the best overall performance. Recall and Precision are somewhat contradictory and are dependent on the choice of the operating point on the PR curve and are indicative only for the chosen point.

Table 13. The 10-fold CV AUC-PR results for the Harvey dataset

Fold	BERT-Linear	Proposed Classifier
Fold 1	0.928	0.9343
Fold 2	0.908	0.9128
Fold 3	0.8399	0.9766
Fold 4	0.9758	0.9918
Fold 5	0.9321	0.9743
Fold 6	0.9735	0.9527
Fold 7	0.9752	0.9853
Fold 8	0.8423	0.9572
Fold 9	0.9177	0.9671
Fold 10	0.9586	0.9689
Average	0.9251	0.9621
Stdev	0.05	0.02
$T-test$	0.047	

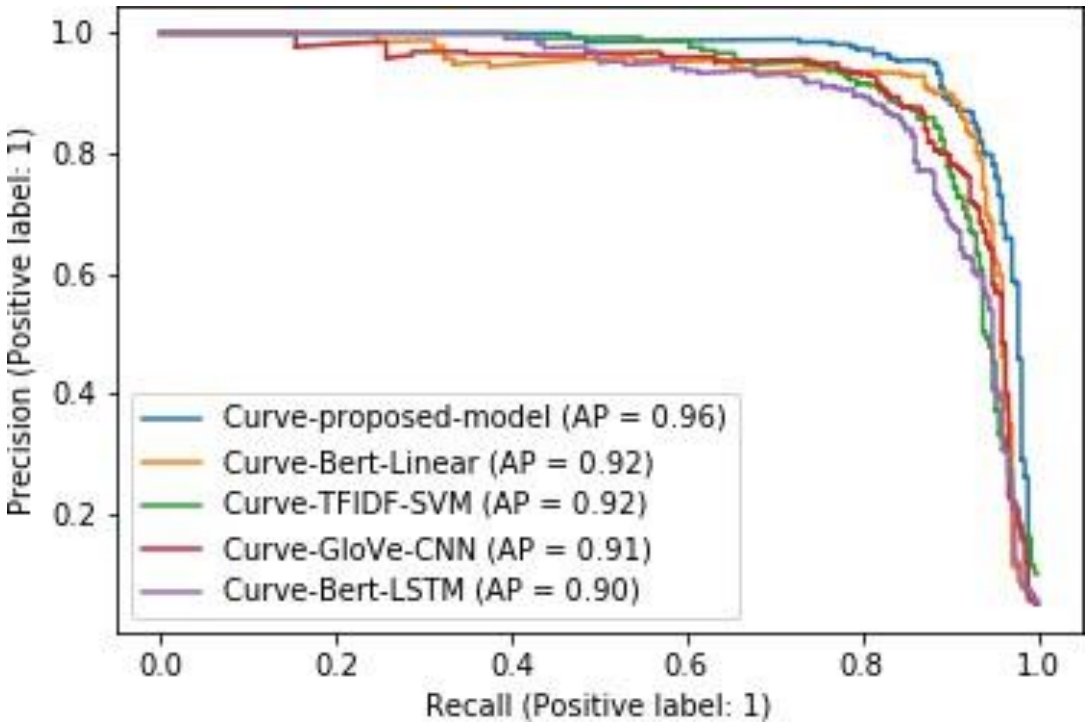

Figure 30. AUC-PR curves on the Harvey dataset

A higher Recall typically means a lower Precision rate, and vice versa. The F1 score is a weighted average of Recall and Precision at the chosen point, which is also a good indication of the overall model performance. The difference between the integrated classifier and the BERT-Linear model is that the proposed integrated classifier adds high-level features derived from analyzing the logical structure of the problem. In terms of the AUC-PR metric, the proposed model outperformed the BERT-Linear classifier by more than 3%. The *t*-test between the 10 testing fold results revealed a statistically significant difference (*p*-values below 0.05) between the AUC-PR results. Overall, both BERT-Linear and the proposed model outperformed the other competing methods in terms of AUC-PR. The confusion matrices (given in Figures 31, 32, 33, 34, and 35) reveal that the GloVe- CNN and TFIDF-SVM models produce a high number of false negatives (e.g., 61, 42, and 40 false negatives given by the GloVe-CNN, TFIDF-SVM, and BERT-LSTM models, respectively). In contrast, BERT-Linear and the proposed classifier reported considerably fewer false negatives (22 and 30 given by the end-to-end fine-tuned BERT and the proposed classifier, respectively).

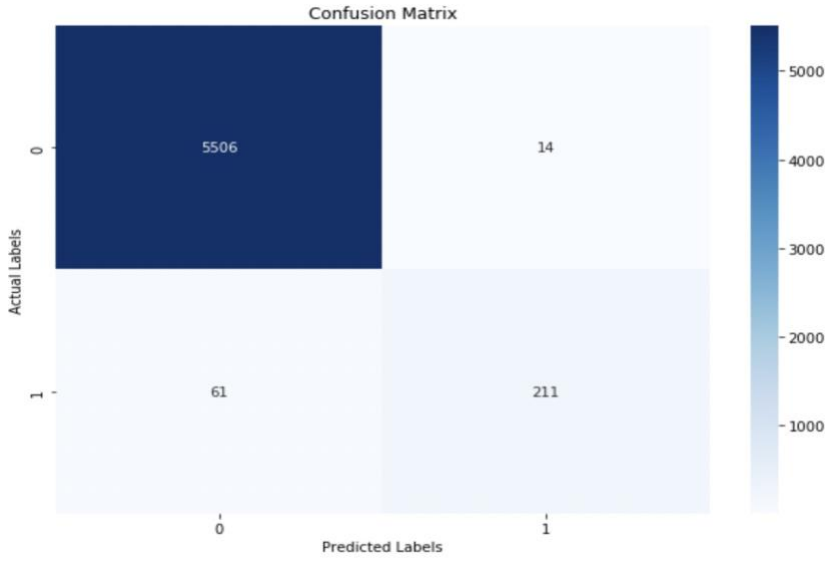

Figure 31. Confusion matrix for GloVe+CNN model on Harvey data set

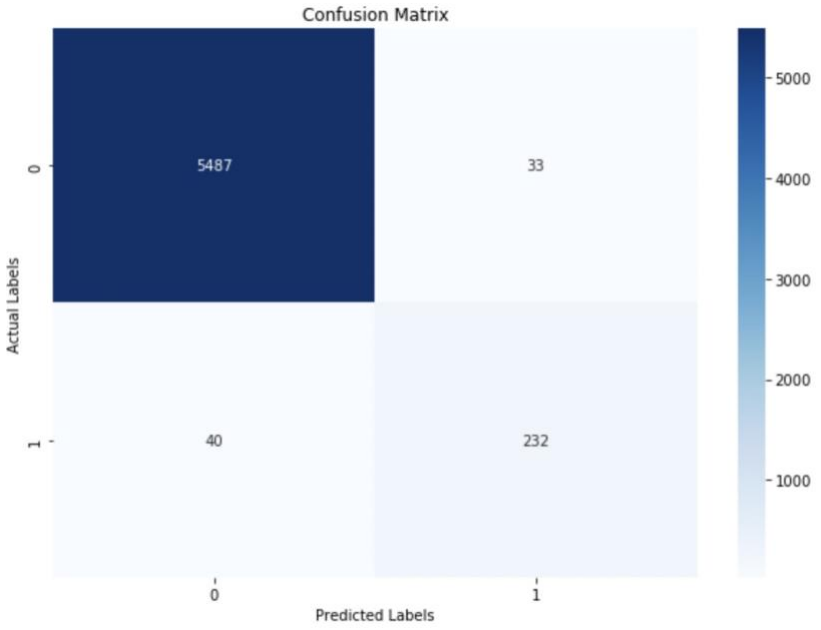

Figure 32. Confusion matrix for TFIDF+SVM model on Harvey data set

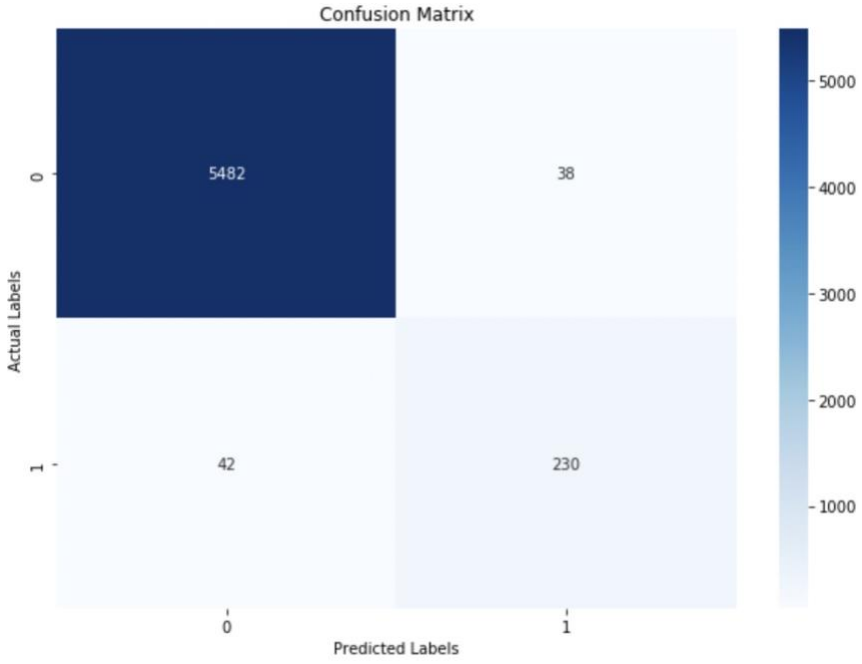

Figure 33. Confusion matrix for BERT+LSTM model on Harvey data set

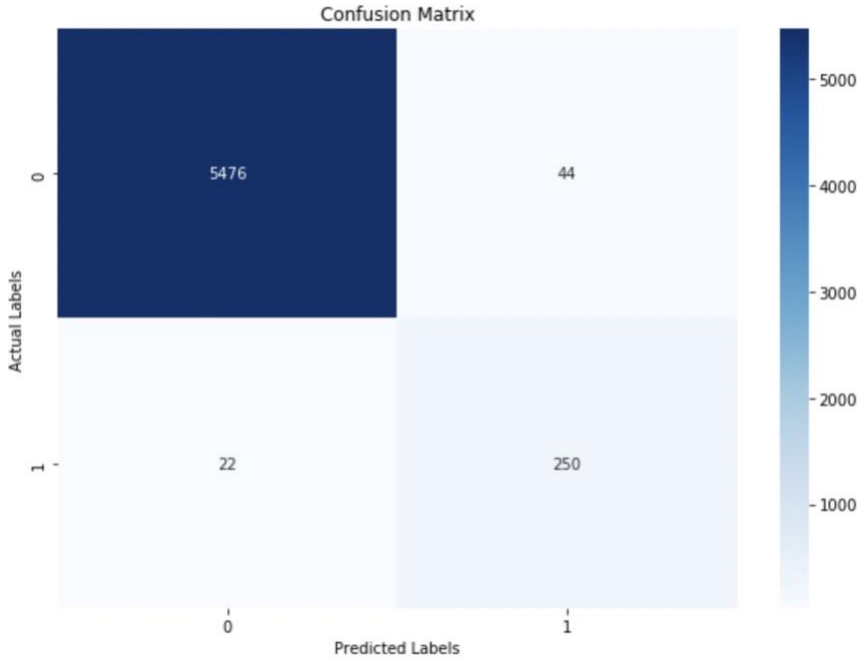

Figure 34. Confusion matrix for BERT+Linear model on Harvey data set

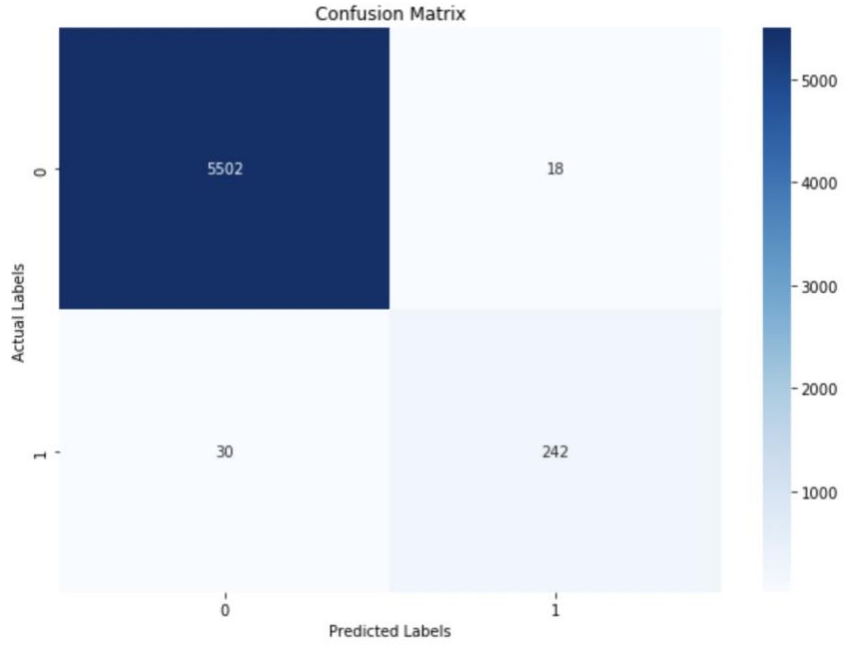

Figure 35. Confusion matrix for the proposed integrated model on Harvey data set

4.1.3.2 Cross-validation on Ian/Ida dataset

The 10-fold cross-validation results on Ian/Ida are also shown in Table 14. A paired *t*-test was conducted to compare the AUC-PR scores obtained by the BERT+linear model and those by the proposed integrated model. Table 15 reports the AUC-PR results for each testing fold in the 10-fold cross-validation for the Harvey dataset. The prediction results from the 10 testing folds were concatenated, and the AUC-PR curves were plotted for the different models, as shown in Figure 36. The confusion matrices obtained by the different models on the Ian/Ida data set are given in Figures 37, 38, 39, 40, and 41. The experiments on the Ian/Ida dataset provided results similar to those from the Hurricane Harvey dataset. Both the BERT+Linear classifier and the proposed model outperformed all other competing methods in terms of AUC-PR and F1. The proposed model showed a slight improvement of 2% in AUC-PR over the BERT-Linear classifier. This difference was statistically significant, as shown by the *t*-test in Table 15. These results indicate that the proposed classifier offers more balanced classification outcomes.

Table 14. 10-fold cross-validation results on Ian/Ida data set

Classifier	AUC-PR	F1-score	Recall	Precision
TFIDF+SVM [34]	0.9263 ± 0.02	0.8424 ± 0.04	0.7873 ± 0.06	0.9105 ± 0.04
BERT+LSTM [148]	0.9206 ± 0.03	0.7905 ± 0.07	0.7557 ± 0.12	0.8494 ± 0.10
GloVe+CNN [34]	0.9291 ± 0.03	0.8485 ± 0.05	0.7996 ± 0.07	0.9212 ± 0.070
BERT+Linear	0.9431 ± 0.03	0.8774 ± 0.04	0.8891 ± 0.06	0.8693 ± 0.05
Integrated classifier	0.9614 ± 0.01	0.9047 ± 0.04	0.9114 ± 0.06	0.9013 ± 0.05

Table 15. The 10-fold CV AUC-PR results for the Ian/Ida dataset

Fold	BERT-Linear	Proposed Classifier
Fold 1	0.9158	0.9404
Fold 2	0.8802	0.9669
Fold 3	0.9820	0.9922
Fold 4	0.9258	0.9465
Fold 5	0.9573	0.9592
Fold 6	0.9891	0.9959
Fold 7	0.9473	0.9526
Fold 8	0.9519	0.9596
Fold 9	0.9610	0.9624
Fold 10	0.9202	0.9385
Average	0.9431	0.9614
Stdev	0.03	0.01
$T-test$	0.047	

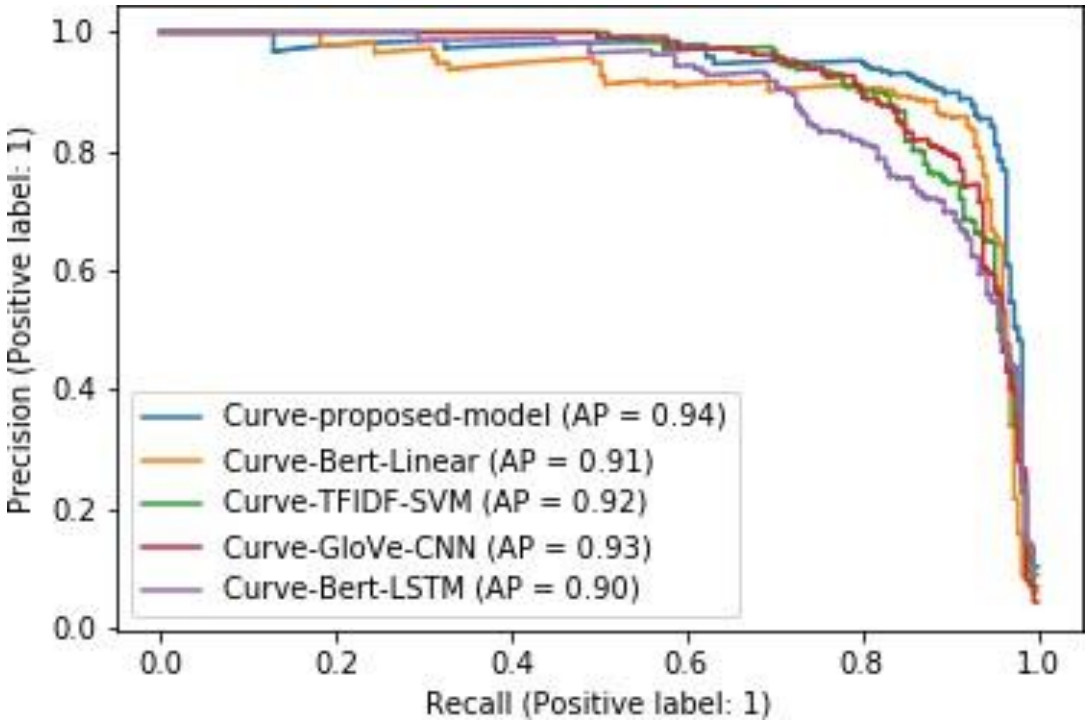

Figure 36. AUC-PR curves on the Ian/Ida dataset

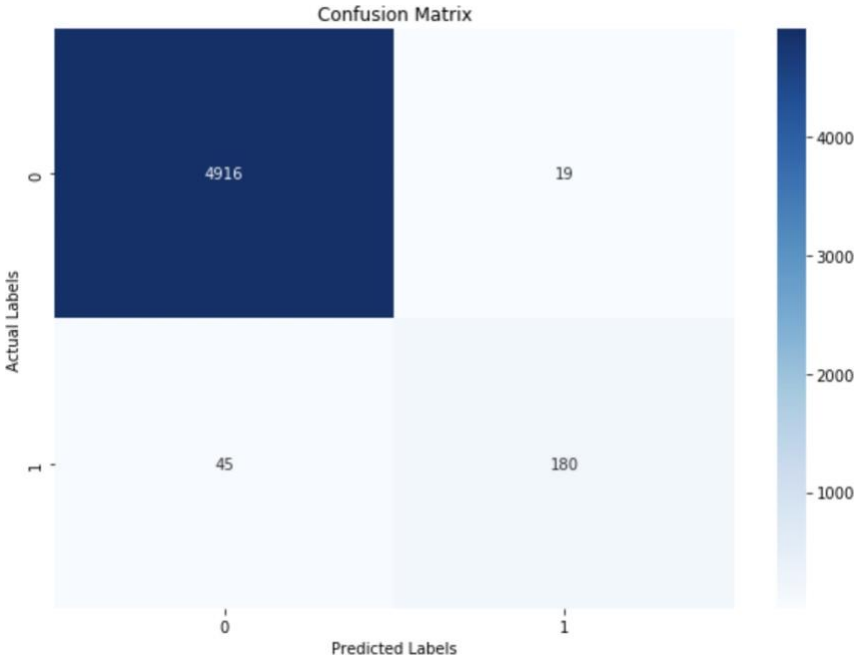

Figure 37. Confusion matrix for GloVe+CNN model on Ian/Ida data set

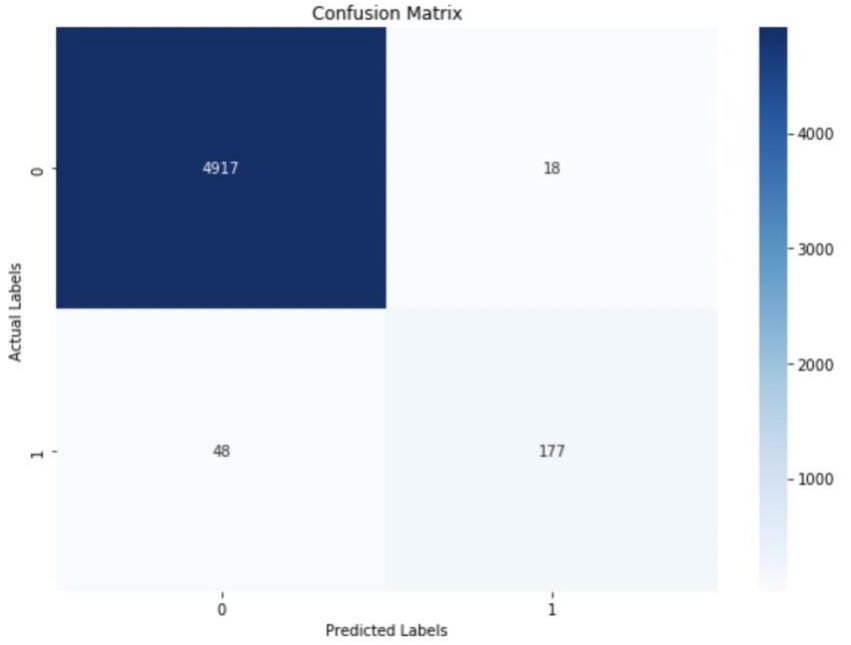

Figure 38. Confusion matrix for TFIDF+SVM model on Ian/Ida data set

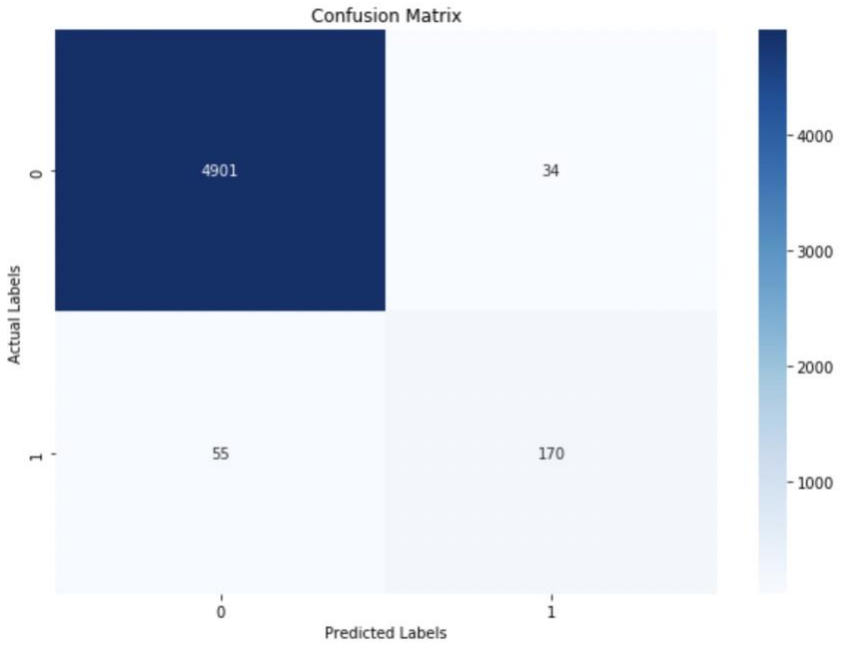

Figure 39. Confusion matrix for BERT+LSTM model on Ian/Ida data set

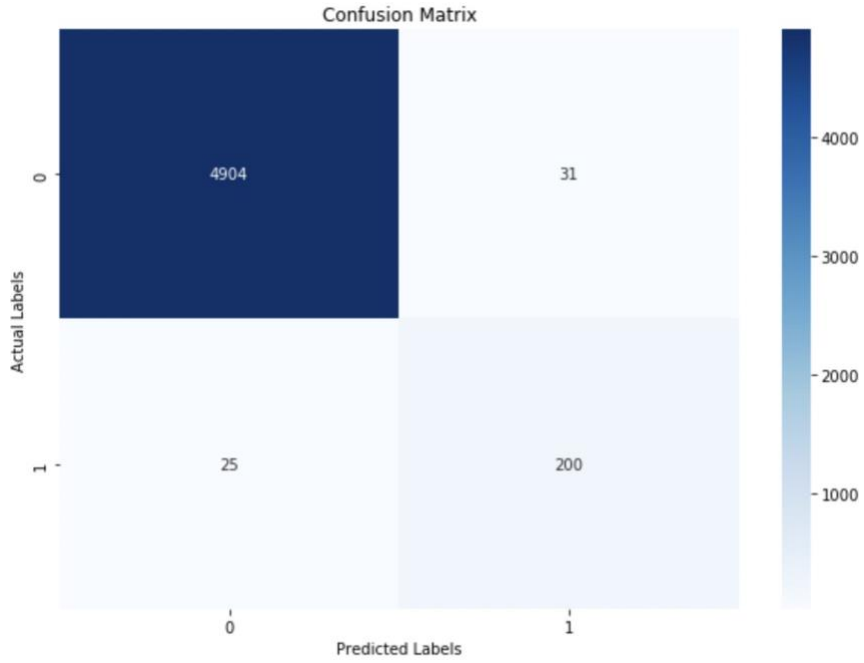

Figure 40. Confusion matrix for BERT+Linear model on Ian/Ida data set

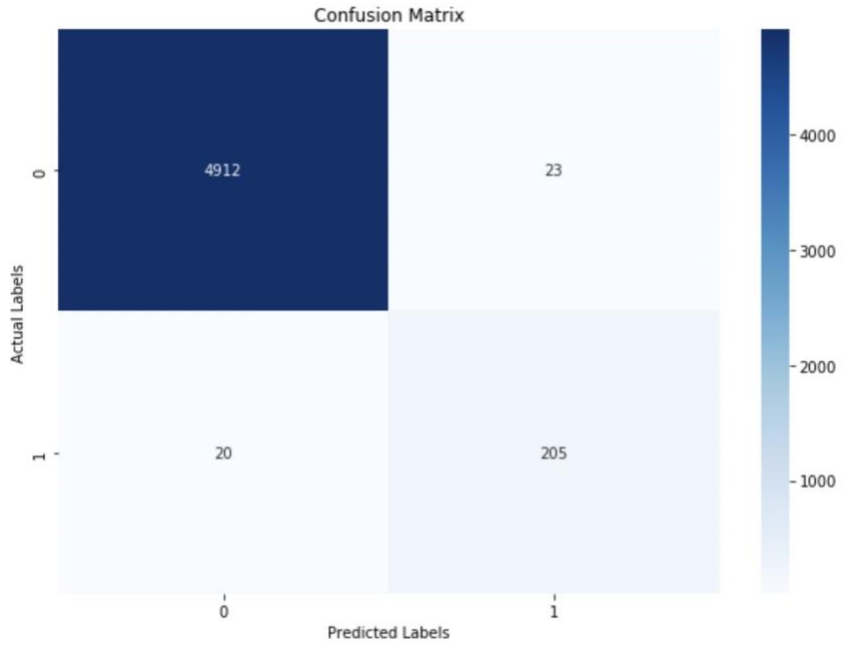

Figure 41. Confusion matrix for the proposed integrated model on Ian/Ida data set

4.1.3.3 Case study for emergency tweets identification

Table 16 illustrates a few prediction examples from Harvey's dataset provided by each classifier. These examples include both correctly classified and incorrectly classified tweets. There are two types of rescue tweets: (1) those including a complete U.S. address of the urgent request (full-address tweet), and (2) those including a broad (fuzzy) location description (partial-address tweet). A typical full-address tweet (e.g., 'There is a family @12423 Meadow Frost Lane, Houston 77044 in dire need of rescue. Thank you! #HoustonStrong #HurricaneHarvey') indicates the location of the rescue request or call for help. However, tweets with partial location descriptions (e.g., 'These are pictures of my uncle's house. He lives in Kingwood. There is no way to see the street sign, but please help! There are kids here too. https://t.co/HzPiqGwuk4' and 'Some people still need rescuing in Cherry Tree Lane in Friendswood. @KHOU #Harvey') provide broader location information without a precise U.S. address but are still valuable for

understanding the situation on the ground and guiding rescue efforts. The first two tweets in Table 16 show typical examples of emergency rescue tweets that were correctly identified by all classifiers. Overall, these tweets are easy to identify and often include complete U.S. addresses and rescue keywords (e.g., 'trapped', 'please help', 'water rescue', etc.). However, it can be noticed that BERT-Linear and the proposed integrated models performed better at detecting full-address tweets compared to GloVe-CNN and TFIDF-SVM (see 3rd, 6th, and 8th tweets in Table 16). There are many similar examples in the dataset.

The proposed model successfully detected tweets with fuzzy and partial addresses, such as '#abc13houston We need help out here. WALLISVILLE RD, PINE TRAILS subdivision Greens Bayous is flooding my neighborhood'. This tweet was misclassified by all other competing methods. The proposed model utilized rescue-related keywords/patterns (e.g., 'We need help', 'is flooding', etc.) as features, which enhanced its ability to detect these tweets. However, there are still some challenging tweets that were missed by all methods, such as 'Some people still need rescuing in Cherry Tree Lane in Friendswood. @KHOU #Harvey'.

Most false negative tweets (i.e., rescue tweets that were incorrectly classified as non-rescue) have broad location descriptions (partial-address tweets). For example, among the 23 false negative tweets by the BERT-Linear classifier in the Harvey dataset, only 5 are full-address tweets. For instance, the rescue tweet 'Some people still need rescuing in Cherry Tree Lane in Friendswood. @KHOU @ #Harvey' has a fuzzy address description and was missed by all methods. Therefore, future research will focus on developing robust classifiers that can detect rescue tweets with fuzzy and partial location descriptions. Many of the false positive tweets (i.e., non-rescue tweets that were incorrectly classified as rescue tweets) included both full-address and partial-

Table 16. Examples of emergency tweet identification by different models

Tweet	True label	Proposed Model	BERT-Linear	TFIDF-SVM	GloVe-CNN	BERT-LSTM
Her address is 618 Regal Street. Houston, TX 77034 and they have been trapped in water through the night. Please help! #HoustonFlood #Harvey https://t.co/ZDtPGngObX	1	1	1	1	1	1
@SheriffEd_HCSO My family is still waiting for a water rescue in the attic, they can't get to roof! 6606 Reamer St Houston, TX 77074	1	1	1	1	1	1
My friends apartments are flooding bad can y'all find somebody to come get them and take them to cypress station from 11800 grant rd 77429,	1	1	1	0	0	1
Some people still need rescuing in Cherry Tree Lane in friendswood. @KHOU @ #Harvey	1	0	0	0	0	0
someone in the 4600 block of Huisache check on elderly lady on corner of Ave B + Huisache ? Not answering phone. #bellaire #harvey #houston	1	1	0	1	1	0
#abc13houston We need help out here. WALLISVILLE RD, PINE TRAILS subdivision Greens bayous is flooding my neighborhood	1	1	0	0	0	0
Any one with a boat near N. Eldridge Parkway we have someone here who needs to check on his parents on enclave pkwy. #ABC13 #fox26	1	1	0	1	0	0
@MsCoCoDominguez My family is currently no flooding and our house is at 2819 barrow creek lane 77089 and we want the evacuation route pls	0	1	1	0	0	0

address tweets. Most of these were tweets offering help (e.g., ‘If anyone near Peach Creek needs help, message me. My dad and brother are out there rescuing with a boat’), sharing rescue updates (e.g., ‘@HCSOTexas @houstonpo- lice The elderly couple at 6603 Mariner Square, Richmond, TX 77407 has been rescued. Thank you, first responders! #HoustonStrong’), or posting advertisements (e.g., ‘Pet-Friendly #Harvey Shelter! TheMET Church, 13000 Jones Rd., #Houston 77070. Thank you @TheMETChurch & @Fox26Houston!’). Compared to other competing models, the proposed classifier significantly reduced the number of false positives, demonstrating a greater ability to capture the contextual details of the tweets.

4.2 RESULTS FOR THE RELIABILITY ASSESSMENT PROBLEM

This section reports the results related to the reliability assessment of rescue tweets posted during natural disasters.

4.2.1 Experimental Setup

4.2.1.1 Data sets

In this part, 472 rescue tweets posted during Hurricane Harvey were labeled by reliability, forming a total of 141 rescue claims. Among the 141 claims collected from Hurricane Harvey, 87 were located in a FEMA-impacted zone and categorized as high-reliability rescue claims, while 54 rescue claims were categorized as low-reliability. The entire dataset was labeled by the author of this dissertation. The class distribution is provided by Table 17. Due to the small size of the data, 5-fold cross-validation was used to select and evaluate the machine learning models.

Table 17. Class distribution of the labeled Harvey data set (by reliability)

	High-reliability	Low-reliability
Tweets	361	111
Rescue claims	87	54

For each annotated tweet, post-related and user-related features were collected as illustrated in Table 18. These features are used for training the machine learning models.

Table 18. Collected Tweets' attributes

Attribute	Dimension
Username	User-related
Account creation year	User-related
Follower count	User-related
Followee count	User-related
Is verified?	User-related
Geo-tagged	User-related
Profile description	User-related
Location	Tweet-related
Posting time	Tweet-related
Attachment	Tweet-related
Replies count	Tweet-related
Retweet count	Tweet-related
is Retweet or Reply?	Tweet-related
Risk zone	Contextual

4.2.1.2 Hyperparameter optimization for machine learning models

The proposed reliability assessment model was compared to a set of supervised machine learning models, including Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), AdaBoost (ADA), and Logistic regression (LR). A grid search with 5-fold cross-validation was employed to determine the best-performing model configuration for each machine learning

architecture. The hyperparameters search spaces for the machine learning models are reported in Table 19. The models achieving the highest average accuracy over the 5 folds were selected. The optimal hyper- parameter values of the best-performing models are illustrated in Table 20.

Table 19. Hyperparameters space for the machine learning models

ML model	Hyperparameter	Space values
DT	Criterion	gini, entropy
	max_depth	2,4,6,8,10,12
	max_features	3,4,5,6
RF	bootstrap	true, false
	max_features	2,4,6,8,10,12
	n_estimators	3,4,5,6
NB	Alpha	1,2,3,4
LR	C	0.001,0.01,0.1,1,10
	penalty	l1, l2, elasticnet, None
	solver	lbfgs, liblinear, newton-cg, newton-cholesky, saga
ADA	learning_rate	0.001,0.01,0.1,1,10
	n_estimators	l1, l2, elasticnet, None

Table 20. Machine learning models selected by grid search)

Classifier	Data	Selected model
DT	Harvey	(Criterion = “gini” , max_depth = 4, max_features = 3)
RF	Harvey	(bootstrap: “true”, max_features: 2, n_estimators: 100)
ADA	Harvey	(learning_rate: 0.1, n_estimators: 300)
NB	Harvey	(alpha: 1)
LR	Harvey	(C: 0.001, penalty: “l1”, solver: “liblinear”)

4.2.1.3 Experiments outline

To evaluate the effectiveness of the proposed reliability assessment model, the obtained confidence scores by the model were converted into binary labels: (1) high-reliability score and (2) low-reliability score. The proposed model's outcome was compared against those of competing methods. The objective of this experiment is to evaluate the performance of the proposed model in distinguishing between high-reliability rescue claims originating from flooded areas and low- reliability rescue claims originating from non-impacted areas.

4.2.2 Results By the Reliability Assessment Model

The reliability assessment framework involves multiple parameters: (1) source weight, (2) content weight, (3) radius, (4) risk value, and (5) threshold value. The source weight w_u and content weight w_t reflect the significance of the source and content variables in the calculated tweet-level reliability scores, with possible values ranging from 0 to 1. The radius *r*, measured in miles, denotes the size of the area used to count the number of rescue-seeking posts in the vicinity of a given claim. The risk factor is a value between -1 and 1 that was used to adjust the score calculations at the claim level. Finally, the threshold value defines the boundary value used to determine the reliability of an incoming rescue claim. The sigmoid damp factors were fixed at 1 for the proximity sigmoid function and 0.4 for the normalization sigmoid function. A range of values for each input parameter was generated as illustrated in Table 22. An exhaustive search was performed to identify the optimal parameters through all possible combinations of values in the search space. The optimal parameters' values of the reliability assessment model are illustrated in Table 21.

Table 21. Optimal parameters for the scoring model

Parameter	Values
W_{user}/	0.25
$W_{content}$	0.75
Radius	0.3m
Risk Penalty	1
Threshold	0.65

Table 22. Proposed model's search space

Parameter	Values
W_{user}	0,0.25,0.5,0.75,1
$W_{content}$	0,0.25,0.5,0.75,1
Radius	0.3,0.5,1
Risk Penalty	-1,-0.5,0,0.5,1
Threshold	0.4,0.45,0.5,0.55,0.6,0.65,0.7,0.75,0.8,0.85,0.9

The proposed reliability assessment model achieved a macro-accuracy of 0.7872. The accuracy for the positive class (i.e., high-reliability rescue claims located in an impacted zone) was 0.8160, while it was 0.7407 for the negative class (i.e., low-reliability rescue claims located in non-impacted areas). The confusion matrix generated by the proposed scoring model is displayed in Figure 42. The obtained results for each class are also summarized in Table 23, where class 1 refers to the high-reliability claims and class 0 to the low-reliability claims.

Table 23. Results by the reliability assessment model for each reliability class

Data	**Label**	Acc_i	$F1_i$	R_i	P_i	
Harvey	class 1	0.8160	0.8255	0.8160	0.8352	
	class 0	0.7407	0.7272	0.7407	0.7142	
	Macro-Acc					0.7872
	FPR					0.2592
	TPR					0.8160
	TNR					0.7407

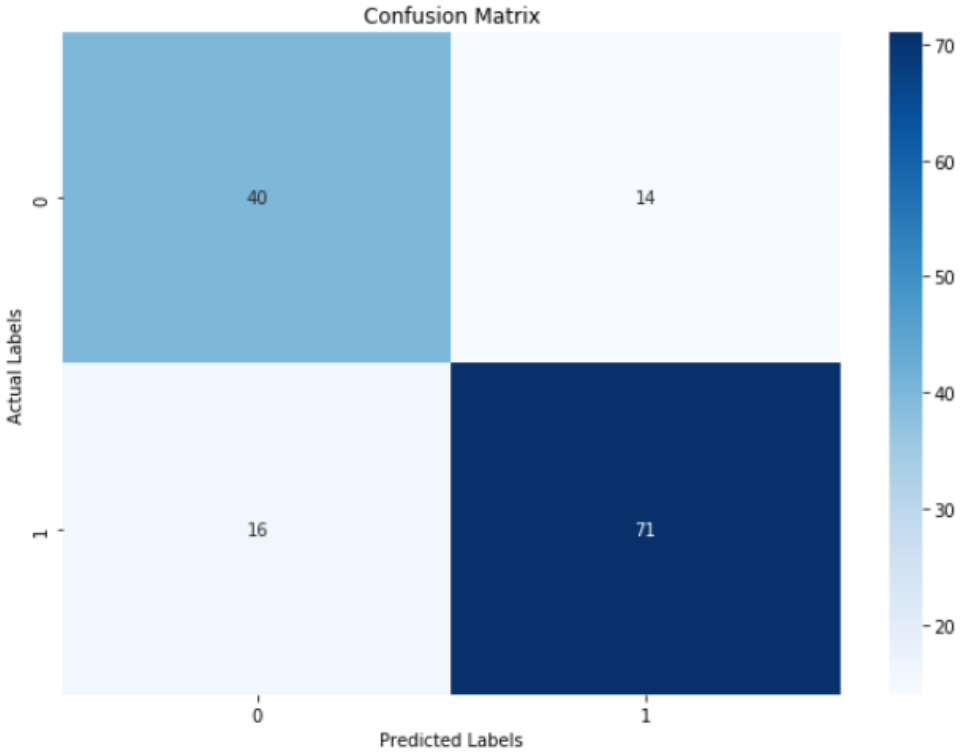

Figure 42. Confusion matrix obtained by the reliability assessment model

The results demonstrate the proposed model's effectiveness in predicting rescue calls originating from impacted and damaged areas within the Houston area.

4.2.3 Comparative Analysis

The proposed reliability assessment model was compared to a set of common supervised learning models, including Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), AdaBoost (ADA), and Logistic Regression (LR), as well as the unsupervised model proposed by Assery et al. [16]. Grid search and 5-fold cross-validation were employed to determine the optimal hyperparameter values for each machine learning model. The predicted values from the selected machine learning models (obtained through 5-fold cross-validation) were then directly compared to the results of the proposed reliability assessment model for all data points. The competing machine learning models make predictions at the tweet level. To convert them to claim level, a majority voting approach was used. For a given rescue claim, if the majority of its related tweets are 'reliable' (i.e., assigned with a 'high-credibility' label), the claim is also labeled as 'high-credibility'. Table 24 presents the comparative analysis results with the competing models. At this stage of the research, the analysis was performed on Harvey's data. In this table, the results are reported in terms of Macro-accuracy (average accuracy), average F1 score, true positive rate (high-credibility categories), true negative rate (low-credibility categories), and FPR (false alarm rate). The proposed two-stage reliability assessment model provides the best accuracy among the competing methods. Assery et al. [16] model provide the worst performance in all metrics. It uses content and user-related features for calculating reliability, which was not sufficient to provide good accuracy on the annotated data proposed by this study. Contextual indicators play a significant role in improving the reliability model's performance compared to the previous Assery et al. [16] model. The best-performing machine learning model among all the selected models, according to the results, is random forest (RF). The proposed two-stage reliability assessment model proposed in this dissertation outperformed RF by a margin of more than 8% in terms of accuracy

and a very large margin in terms of F1 score (from 0.5845 to 0.7764). However, the matching learning model can be further improved by collecting and labeling more data. RF achieves a 100% true positive rate. However, this rate is coming at the expense of the true negative rate which was very low (e.g., 0.22 for RF). Overall, the proposed two-stage reliability assessment model was able to achieve a good balance between true positive and true negative rates compared to the competing models.

Table 24. Comparative analysis on Hurricane Harvey data)

Model	**Macro-Acc**	**Avr F1**	**TPR**	**TNR**	**FPR**	**TP**	**TN**	**FP**	**FN**
DT	0.6424	0.4875	0.9885	0.111	0.8888	86	6	48	1
RF	0.7021	0.5845	1	0.2222	0.7777	87	12	42	0
ADA	0.6453	0.4573	1	0.074	0.9259	87	4	50	0
NB	0.4042	0.3224	0.0459	0.9814	0.0185	4	53	1	83
LR	0.6170	0.3815	1	0	1	87	0	54	0
Assery model [16]	0.3758	0.3182	0.068	0.8703	0.1296	6	47	7	81
Two-stage model	0.7872	0.7764	0.816	0.7407	0.2592	71	40	14	16

CHAPTER 5

DISCUSSION

Social media platforms have emerged as an alternative means of communication during natural disasters. Reports from previous hurricanes indicate an increasing trend of individuals using social media networks, such as Twitter, to share information, request help, and react to emergencies during such events. Developing automated systems to extract actionable information from social media platforms is crucial. Researchers have proposed various computational methods, primarily machine learning methods, to extract useful information that meets disaster responders' needs. Most of these methods focus on the generalized concept of 'Situational awareness'. Identifying actionable messages posted on social media during natural disasters, such as implicit and explicit rescue requests, has received relatively less attention. To fill this gap, this study introduces novel methods for extracting reliable rescue information from Twitter during hurricanes. This chapter presents the key findings and limitations of this research, discusses the practical implications, and outlines future research directions.

5.1 FINDINGS

The first part of this study investigates the problem of identifying actionable emergency rescue tweets during hurricanes. This problem has been explored in a few prior research studies (e.g., [148], [137], and [34]). Researchers have applied supervised machine learning and deep learning methods, which learn textual features directly from raw tweets. This study investigates a novel

approach to improve the proposed rescue messages identification methods by using rescue related features (domain-dependent features) identified through regular expression (*regex*). The proposed features consist of (1) contextual features, (2) ask-for-help features, (3) rescue hashtags, and (4) address/location features. A set of non-rescue features—intended for exclusion—was also identified, including (5) political tweets, (6) offering-help features, (7) commercial tweets, (8) news reports, and (9) situational and rescue updates.

This study presents two models: (1) a logic-based model that uses *regex* features, and (2) an integrated classification model, which employs two types of features: high-level problem-specific features and low-level statistical features derived from a fine-tuned BERT model. The experiments showed promising outcomes of the logic-based approach. The logic-based approach achieved F1 scores above 0.81 for both data set, showing an acceptable balance between the precision of the classifier and recall. This approach does not require training and produces an explainable classification output. Consequently, it can be employed as a quick method to collect rescue tweets as a new disaster unfolds.

This study found that combining the two sets of features (low-level and high-level features) enhanced the model's performance, as measured by the area under the precision-recall curve, achieving a better precision-recall balance. Furthermore, the experimental results demonstrated that both the BERT-Linear classifier and the integrated classifier significantly outperformed competing methods, such as TFIDF-SVM, Glove-CNN, and BERT-LSTM VictimFinder, in retrieving emergency rescue requests from large volumes of social media data. This finding aligns with previous research indicating that transformer models are highly effective in extracting rescue requests [148]. Furthermore, traditional classification models, such as CNN and SVM, are shown to be less effective for this problem and often fail to detect urgent tweets during crises.

BERT-LSTM (VictimFinder) surprisingly underperformed on both datasets compared to BERT-Linear and the proposed model. This could be attributed to the core BERT model used in BERT-LSTM being pretrained on a general corpus, which limits its ability to learn task-specific features. However, the core BERT model in the BERT-Linear classifier was fine-tuned end-to-end on the labeled dataset, resulting in better classification performance. The proposed model also used a fine-tuned BERT model for low-level feature generation, indicating the importance of using labeled examples from newly incoming hurricane events to calibrate the parameters of the transformer model. The area under the precision-recall curve (AUC-PR) metric provides an overall view of a model's ability to balance precision and recall, determining which model offers an optimal balance. In the disaster response context, accurately identifying rescue requests while minimizing false positives is crucial for decision-makers to ensure resources are not wasted on false positive alerts. For both datasets, the AUC-PR metrics indicated that the proposed model performed best among all competing models.

Both disaster response practitioners and researchers have acknowledged that first-hand information shared on social media channels is a valuable source of real-time information that should be integrated into the formal emergency workflow [77]. Nevertheless, the veracity of social media posts, especially those posted by the general public, remains a major concern [144]. The assessment of the reliability of social media messages during natural disasters poses a significant challenge. Calls for help and rescue messages are particularly difficult to verify during disasters. This research is a step forward in addressing this issue. While previous studies have focused on assessing the reliability of social media data in different contexts, such as general news and medical-related messages, the reliability of actionable social media information

in a disaster context has not been addressed. More specifically, none of the previous studies has addressed the reliability of the rescue requests posted on social media during disaster events.

This study introduces a two-stage reliability scoring model to estimate the reliability of hurricane- related rescue messages. The proposed model was quantitatively evaluated on an annotated set of rescue tweets posted during Hurricane Harvey. This study found that the proposed two-stage reliability assessment model was effective in categorizing both high-reliability and low-reliability rescue tweets, where reliability was approximated by the tweets' locations within the FEMA dam- age assessment map. Machine learning models trained on the annotated dataset using user-related, post-related, and context-related features did not perform well in categorizing rescue tweets by reliability. This poor performance could be attributed to the small size of the dataset used for training. The best performing machine learning model was the random forest (RF). RF employs a rule-based approach for classification, making it more efficient with such a small dataset. This study found that the proposed reliability assessment model outperformed the RF by a margin of 8% in accuracy. The proposed reliability assessment model was compared to the unsupervised approach for reliability assessment proposed by Assery et al. [16]. The Assery model provided the worst performance among all models. This study found that the proposed reliability assessment model significantly outperforms the Assery reliability scoring model by very large margins. This could be attributed to the contextual indicators used in the proposed two-stage model that were not used in the previous model. The proposed two-stage reliability assessment model was compared to the unsupervised approach proposed by Assery et al. [16]. The Assery model provided the worst performance among all models. This study found that the proposed two-stage reliability assessment model offers

significant improvements over the Assery model measured by several metrics. Overall, this study demonstrates that the proposed reliability indicators provide better outcomes than those achieved by training machine learning models.

5.2 LIMITATIONS

The integrated classification model proposed in this study was evaluated on two distinct datasets representing different hurricanes. Unlike previous studies that focused on data from Hurricane Harvey, this study includes data from several hurricane events for more robust results. It is important to test the model's performance across a broader range of hurricanes. Furthermore, it remains uncertain how the model would perform for other disaster events, such as fires, earthquakes, and man-made disasters since the proposed model was tested only on hurricane rescue tweets. Each disaster event is unique in terms of locations, people involved, and types of information, which might cause slight variations in how rescue-related social media posts are shared. Investigating the textual characteristics of these messages across a wide range of disaster types requires further investigation.

The proposed classification model uses a supervised machine learning approach. To achieve optimal classification accuracy using a transformer-based classifier on a newly incoming (unseen) event, it still requires labeled data from the event to fine-tune the model. Relying solely on historical data is not sufficient. Data annotation is expensive and time-consuming. Therefore, labeling data with every new event is not a practical approach in humanitarian relief scenarios. Furthermore, conducting a cross-domain evaluation of the proposed model is crucial to assess the effectiveness of the model in predicting unseen tweets outside of the specific hurricane on which it was trained. While the integrated classifier demonstrated effectiveness in identifying rescue

requests with full addresses, it still requires improvement in identifying emergency rescue requests with partial location descriptions.

The proposed two-stage reliability assessment model was evaluated on a relatively small data set, which presents a primary limitation of this study. A significant challenge is the high cost of data collection, as human resources are required to label social media messages. At this stage, the proposed model was evaluated using rescue messages collected during hurricane Harvey. The applicability of the proposed reliability assessment model to other event types (e.g., fires, earthquakes) and information types (e.g., infrastructure damage, urgent needs) needs further investigation. Reliability was approximated by the location of the tweet in the FEMA damage assessment map. The results apply to the defined criteria for reliability. However, the definition of reliability should be refined by involving domain experts and establishing new annotation guidelines accordingly.

5.3 IMPLICATION(S)

This research contributes to the crisis informatics body of knowledge and disaster management practice by developing novel models for identifying rescue messages on Twitter during hurricanes and assessing their reliability. The key contributions of this research are summarized as follows:

1. This dissertation introduced two annotated datasets derived from different hurricane events. These data sets can be used in future studies focused on identifying rescue messages from social media during hurricanes. Additionally, this dissertation proposed a data set of rescue tweets annotated by reliability. Despite the limited size of the dataset, there is potential for expansion in future research. The annotation procedure that was employed to label rescue messages can be adopted in future

research investigating the reliability of social media data.

2. This dissertation developed a novel classification framework to identify rescue messages on social media platforms, combining different domain-specific rule-based features with textual features learned by the pretrained language model (BERT) to improve the robustness of the classification model. The rescue features developed in this research showed a good predictive performance in identifying rescue messages. The proposed model achieved a good balance between false negatives and false positives, which is of practical importance for decision- makers and first responders in allocating resources during emergency events.
3. This dissertation developed a credibility assessment framework for evaluating "actionable" rescue information posted on Twitter during hurricane events. The proposed model was evaluated using rescue tweets collected during Hurricane Harvey. This research offered useful insights into these rescue messages from a reliability assessment perspective.

To the best of the author's knowledge, no prior research has examined the reliability of actionable social media information during disasters, such as rescue requests. The proposed model carries a practical implication. It offers analysts an explainable reliability assessment tool to assess posted rescue messages based on several assessment dimensions. This would help analysts and decision-makers to alleviate the information overload problem during disasters. Furthermore, the outcome of the reliability assessment model can assist analysts and decision-makers in building a meaningful operational picture of the situation and enhance 'situational awareness' through reliable information.

From a theoretical perspective, the proposed rescue features and the reliability indicators for building the credibility scoring model provide useful insights to researchers about the characteristics of these messages. Although transformer-based models have made significant advancements in a wide variety of applications, this study found that integrating problem-specific features can further enhance their performance. The findings of this research can be used to build more efficient solutions for extracting useful information from social media platforms.

5.4 RECOMMENDATIONS FOR FUTURE RESEARCH

Research on the reliability of social media information in the context of natural disasters is largely unexplored. The present research represents a step forward in this area. Potential future research directions can be summarized as follows. To overcome the data annotation cost issue, several techniques, including few-shot learning and semi-supervised learning, can be utilized to improve model performance with only a few labeled data samples. Experimental results showed that location information in tweets is crucial but often challenging to extract, as it is not always provided in formal English and may contain grammatical errors, spelling mistakes, abbreviations, etc. Future work includes improving location information extraction from social media posts and extending the emergency scenarios to include wildfires, shootings, and earthquakes. Future research can also include qualitative studies, such as interviews with disaster response practitioners, to enhance understanding of how professionals perceive and evaluate the reliability of social media rescue messages. Such studies can help refine and enhance the proposed automatic tools, as well as develop more accurate annotation schemes for the problem.

REFERENCES

[1] M. Abavisani, L. Wu, S. Hu, J. Tetreault, and A. Jaimes, "Multimodal categorization of crisis events in social media," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 679–14 689.

[2] M.-A. Abbasi and H. Liu, "Measuring user credibility in social media", in *Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA, April 2-5, 2013. Proceedings 6*, Springer, 2013, pp. 441–448.

[3] B. Abu-Salih, P. Wongthongtham, K. Y. Chan, and D. Zhu, "Credsat: Credibility ranking of users in big social data incorporating semantic analysis and temporal factor", *Journal of Information Science*, vol. 45, no. 2, pp. 259–280, 2019.

[4] A. Aipe, N. Mukuntha, A. Ekbal, and S. Kurohashi, "Deep learning approach towards multi-label classification of crisis related tweets", in *Proceedings of the 15th ISCRAM Conference*, 2018.

[5] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework", in *Proceedings of the 25rd ACM SIGKDD Inter- national Conference on Knowledge Discovery and Data Mining*, 2019.

[6] S. Aladhadh, X. Zhang, and M. Sanderson, "Tweet author location impacts on tweet credibility", in *Proceedings of the 19th Australasian Document Computing Symposium*, 2014, pp. 73–76.

[7] F. Alam, F. Ofli, and M. Imran, “Crisismmd: Multimodal twitter datasets from natural disasters”, in *Proceedings of the international AAAI conference on web and social media*, vol. 12, 2018.

[8] F. Alam, F. Ofli, and M. Imran, “Processing social media images by combining human and machine computing during crises”, *International Journal of Human-Computer Interaction*, vol. 34, no. 4, pp. 311–327, 2018.

[9] F. Alam, U. Qazi, M. Imran, and F. Ofli, “Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks”, in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 933–942.

[10] A. A. AlMansour and C. S. Iliopoulos, “Using arabic microblogs features in determining credibility”, in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 2015, pp. 1212–1219.

[11] M. Alrubaian, M. Al-Qurishi, M. M. Hassan, and A. Alamri, “A credibility analysis system for assessing information on twitter”, *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 661–674, 2016.

[12] M. AlRubaian, M. Al-Qurishi, M. Al-Rakhami, M. M. Hassan, and A. Alamri, “Credfinder: A real-time tweets credibility assessing system”, in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2016, pp. 1406–1409.

[13] A. Alshehri and S. Alahamri, “An ensemble learning for detecting situational awareness tweets during environmental hazards”, in *2019 IEEE International Systems Conference (SysCon)*, IEEE, 2019, pp. 1–8.

[14] S. Andrews, T. Day, K. Domdouzis, L. Hirsch, R. Lefticaru, and C. Orphanides, “Analyzing crowd-sourced information and social media for crisis management”, *Application of Social Media in Crisis Management: Advanced Sciences and Technologies for Security Applications*, pp. 77–96, 2017.

[15] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, “Tweedr: Mining twitter to inform disaster response”, in *ISCRAM*, Citeseer, 2014, pp. 269–272.

[16] N. Assery, Y. Xiaohong, Q. Xiuli, R. Kaushik, and S. Almalki, “Evaluating disaster-related tweet credibility using content-based and user-based features”, *Information Discovery and Delivery*, vol. 50, no. 1, pp. 45–53, 2022.

[17] N. D. Association, *Natural hazards | hurricanes*, https://www.n-d-a.org/hurricane. php, Accessed: 2021-06-06, 2021.

[18] D. Baer, *As sandy became #sandy emergency services got social*, https://www.fastcompany. com/3002837/sandy-became-sandy-emergency-services-got-social, Accessed: 2020-09-10, 2012.

[19] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman, “Challenges of computational verification in social multimedia”, in *Proceedings of the 23rd international conference on world wide web*, 2014, pp. 743–748.

[20] C. Buckner, “Empiricism without magic: Transformational abstraction in deep convolutional neural networks”, *Synthese*, vol. 195, no. 12, pp. 5339–5372, 2018.

[21] K. R. Canini, B. Suh, and P. L. Pirolli, “Finding credible information sources in social networks based on content and social structure”, in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, IEEE, 2011, pp. 1–8.

[22] C. Caragea, A. Silvescu, and A. H. Tapia, “Identifying informative messages in disaster events using convolutional neural networks,” in *International conference on information systems for crisis response and management*, 2016, pp. 137–147.

[23] Y. Cardinale, I. Dongo, G. Robayo, D. Cabeza, A. Aguilera, and S. Medina, “Tcreo: A twitter credibility analysis framework,” *IEEE Access*, vol. 9, pp. 32 498–32 516, 2021.

[24] C. Cea, *Dataset for supervised bot detection on twitter*, version 1.0, Zenodo, Oct. 2021.

DOI: 10.5281/zenodo.5574403. [Online]. Available: https://doi.org/10.5281/zenodo.5574403.

[25] X. H. Center, *How to get the blue checkmark on x*, https://help.twitter.com/en/managing- your- account/ legacy- verification- policy, Accessed: 2023-12-23, 2023.

[26] V. Chamola, V. Hassija, S. Gupta, A. Goyal, M. Guizani, and B. Sikdar, “Disaster and pan- demic management using machine learning: A survey”, *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 16 047–16 071, 2020.

[27] M. Choudhary, S. S. Chouhan, E. S. Pilli, and S. K. Vipparthi, "Berconvonet: A deep learning framework for fake news classification", *Applied Soft Computing*, vol. 110, p. 107 614, 2021.

[28] E. Ciceri, R. Fedorov, E. Umuhoza, M. Brambilla, and P. Fraternali, "Assessing online media content trustworthiness, relevance and influence: An introductory survey", in *KDWeb*, 2015, pp. 29–40.

[29] J. Coche, J. Kropczynski, A. Montarnal, A. Tapia, and F. Benaben, "Actionability in a situation awareness world: Implications for social media processing system design", in *ISCRAM 2021-18th International conference on Information Systems for Crisis Response and Management*, 2021, p–994.

[30] C. Cortes and V. Vapnik, "Support-vector networks", Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[31] M. Coyne, *Service assessment august-september 2021 hurricane ida*, https : / / www . weather . gov / media / publications / assessments / Hurricane _ Ida _ Service _ Assessment.pdf, Accessed: 2023-08-28, 2023.

[32] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[33] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves", in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[34] A. Devaraj, D. Murthy, and A. Dontula, “Machine-learning methods for identifying social media-based requests for urgent help during hurricanes”, *International Journal of Disaster Risk Reduction*, vol. 51, p. 101 757, 2020.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint arXiv:1810.04805*, 2018.

[36] S. Doroudi, “On the paradigms of learning analytics: Machine learning meets epistemology”, *Computers and Education: Artificial Intelligence*, vol. 6, p. 100 192, 2024.

[37] M. Eriksson, “Lessons for crisis communication on social media: A systematic review of what research tells the practice”, *International Journal of Strategic Communication*, vol. 12, no. 5, pp. 526–551, 2018.

[38] D. Fallis, “On verifying the accuracy of information: Philosophical perspectives”, *Library trends*, vol. 52, no. 3, pp. 463–487, 2004.

[39] C. Fan, F. Wu, and A. Mostafavi, “A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters”, *IEEE Access*, vol. 8, pp. 10 478–10 490, 2020.

[40] FEMA, *FEMA*, https://www.fema.gov/.

[41] FEMA, *Fema historical geospatial damage assessment database*, https://www.arcgis.com/home/item.html?id=e56e3d16fa144684bd532edbf298042c, Accessed: 2024-2-15.

[42] C. Garcia, G. Rabadi, D. Abujaber, and M. Seck, "Supporting humanitarian crisis decision making with reliable intelligence derived from social media using AI", Journal of Homeland Security and Emergency Management, vol. 20, no. 2, pp. 97–131, 2023.

[43] A. K. Gautam, L. Misra, A. Kumar, K. Misra, S. Aggarwal, and R. R. Shah, "Multimodal analysis of disaster tweets", in *2019 IEEE Fifth international conference on multimedia big data (BigMM)*, IEEE, 2019, pp. 94–103.

[44] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: Real-time credibility assessment of content on twitter", in *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings 6*, Springer, 2014, pp. 228–243.

[45] M. Habdank, N. Rodehutskors, and R. Koch, "Relevancy assessment of tweets using supervised learning techniques: Mining emergency related tweets for automated relevancy classification", in *2017 4th International conference on information and communication technologies for disaster management (ICT-DM)*, IEEE, 2017, pp. 1–8.

[46] S. E. Halse, A. Tapia, A. Squicciarini, and C. Caragea, “An emotional step toward automated trust detection in crisis social media,” *Information, communication & society*, vol. 21, no. 2, pp. 288–305, 2018.

[47] Y. Han, S. Karunasekera, and C. Leckie, “Graph neural networks with continual learning for fake news detection from social media”, *arXiv preprint arXiv:2007.03316*, 2020.

[48] A. Hanchey, A. Schnall, T. Bayleyegn, S. Jiva, A. Khan, V. Siegel, R. Funk, and E. Svendsen, “Notes from the field: Deaths related to hurricane ida reported by media—nine states, august 29–september 9, 2021”, *Morbidity and Mortality Weekly Report*, vol. 70, no. 39, p. 1385, 2021.

[49] H. Hao and Y. Wang, “Leveraging multimodal social media data for rapid disaster damage assessment”, *International Journal of Disaster Risk Reduction*, vol. 51, p. 101 760, 2020.

[50] M. van den Homberg, R. Monné, and M. Spruit, “Bridging the information gap of disaster responders by optimizing data selection using cost and quality”, *Computers & geosciences*, vol. 120, pp. 60–72, 2018.

[51] J. Huang, W. Khallouli, G. Rabadi, and M. Seck, “Intelligent Agent for Hurricane Emergency Identification and Text Information Extraction from Streaming Social Media Big Data”, *International Journal of Critical Infrastructures*, vol 19, no. 2, pp. 124-139, 2023.

[52] J. Huang, W. Khallouli, G. Rabadi, and M. Seck, “Intelligent agent for hurricane emergency identification and text information extraction from streaming social media big data”, *International Journal of Critical Infrastructures*, vol. 19, no. 2, pp. 124–139, 2023.

[53] K. Hunt, P. Agarwal, and J. Zhuang, “Monitoring misinformation on twitter during crisis events: A machine learning approach”, *Risk analysis*, vol. 42, no. 8, pp. 1728–1748, 2022.

[54] K. Hunt, B. Wang, and J. Zhuang, “Misinformation debunking and cross-platform information sharing through twitter during hurricanes Harvey and Irma: A case study on shelters and ID checks”, *Natural Hazards*, vol. 103, no. 1, pp. 861–883, 2020.

[55] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, “AIDR: Artificial intelligence for disaster response”, in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 159–162.

[56] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, “Practical extraction of disaster- relevant information from social media”, in *Proceedings of the 22nd international conference on world wide web*, 2013, pp. 1021–1024.

[57] M. Imran, F. Ofli, D. Caragea, and A. Torralba, “Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions”, *Information Processing & Management*, vol 57, no. 5, pp. 102-261, 2020.

[58] E. Jaho, E. Tzoannos, A. Papadopoulos, and N. Sarris, “Alethiometer: A

framework for assessing trustworthiness and content validity in social media", in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 749–752.

[59] T. Jose and S. S. Babu, "Detecting spammers on social network through clustering technique", *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–15, 2019.

[60] M. Y. Kabir and S. Madria, "A deep learning approach for tweet classification and rescue scheduling for effective disaster management", in *Proceedings of the 27th ACM SIGSPA- TIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 269–278.

[61] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach", *Multimedia tools and applications*, vol. 80, no. 8, pp. 11 765–11 788, 2021.

[62] E. Karam, W. Hussein, and T. F. Gharib, "Integrating location and textual information for detecting affected people in a crisis", *Social Network Analysis and Mining*, vol. 11, pp. 1– 12, 2021.

[63] R. S. Kevin Sullivan and E. Wax-Thibodeaux, *More than 30,000 people expected in shelters as extent of Harvey's blow comes into chilling focus*, https://www.washingtonpost.com/politics/full-extent-of-harveys-aftermath-starts-to-come-into-chilling-focus/2017/08/27/1b2b184a-8b56-11e7-8df5-c2e5cf46c1e2_story.html, Accessed: 2023-08-28, 2017.

[64] Y. Kim, *Convolutional neural networks for sentence classification*, 2014. arXiv:1408.5882 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1408.5882.

[65] K. Kirasich, T. Smith, and B. Sadler, "Random forest vs logistic regression: binary classification for heterogeneous datasets", *SMU Data Science Review*, vol. 1, no. 3, p. 9, 2018.

[66] R. Koshy and S. Elango, "Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model", *Neural Computing and Applications*, pp. 1–21, 2022.

[67] Z. Kou, L. Shang, Y. Zhang, and D. Wang, "Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection", *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. GROUP, pp. 1–25, 2022.

[68] S. Krishnan and M. Chen, "Identifying tweets with fake news", in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, IEEE, 2018, pp. 460–464.

[69] A. Kumar and J. P. Singh, "Location reference identification from tweets during emergencies: A deep learning approach," *International journal of disaster risk reduction*, vol. 33, pp. 365–375, 2019.

[70] A. Kumar, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "A deep multi-modal neural network for informative twitter content classification during emergencies", *Annals of Operations Research*, pp. 1–32, 2020.

[71] A. Kumar, J. P. Singh, and N. P. Rana, “Authenticity of geo-location and place name in tweets”, 2017.

[72] A. Kumar, J. P. Singh, N. P. Rana, and Y. K. Dwivedi, “Multi-channel convolutional neural network for the identification of eyewitness tweets of disaster”, *Information Systems Fron- tiers*, pp. 1–16, 2022.

[73] S. Kumar, G. Barbier, M. Abbasi, and H. Liu, “Tweettracker: An analysis tool for humanitarian and disaster relief”, in *Proceedings of the international aaai conference on web and social media*, vol. 5, 2011, pp. 661–662.

[74] R. Kumari and A. Ekbal, “Amfb: Attention based multimodal factorized bilinear pool- ing for multimodal fake news detection”, *Expert Systems with Applications*, vol. 184, p. 115 412, 2021.

[75] C. Kyrkou, P. Kolios, T. Theocharides, and M. Polycarpou, “Machine learning for emergency management: A survey and future outlook”, *Proceedings of the IEEE*, vol. 111, no. 1, pp. 19–41, 2022.

[76] K. A. Lachlan, P. R. Spence, X. Lin, K. Najarian, and M. Del Greco, “Social media and crisis management: CERC, search strategies, and twitter content”, *Computers in Human Behavior*, vol. 54, pp. 647–652, 2016.

[77] H. Li, D. Caragea, and C. Caragea, “Combining self-training with deep learning for disaster tweet classification”, in *The 18th international conference on information systems for crisis response and management (ISCRAM 2021)*, 2021.

[78] R. Li and A. Suh, “Factors influencing information credibility on social media platforms: Evidence from Facebook pages”, *Procedia computer science*, vol. 72, pp. 314–328, 2015.

[79] X. Li, P. Lu, L. Hu, X. Wang, and L. Lu, “A novel self-learning semi-supervised deep learning network to detect fake news on social media”, *Multimedia tools and applications*, vol. 81, no. 14, pp. 19 341–19 349, 2022.

[80] Y. C. Lin, F. Khan, S. F. Jenkins, and D. Lallemant, “Filling the disaster data gap: Lessons from cataloging Singapore’s past disasters”, *International Journal of Disaster Risk Science*, vol. 12, pp. 188–204, 2021.

[81] V. Linardos, M. Drakaki, P. Tzionas, and Y. L. Karnavas, “Machine learning in disaster management: Recent developments in methods and applications”, *Machine Learning and Knowledge Extraction*, vol. 4, no. 2, 2022.

[82] B. Lisa, A. Laura, H. Andrew, D. Sandy, and B. Jack, “National hurricane center tropical cyclone report: Hurricane Ian”, *National Hurricane Center*, 2023.

[83] J. Liu, T. Singhal, L. T. Blessing, K. L. Wood, and K. H. Lim, “Crisisbert: A robust transformer for crisis classification and contextual crisis embedding”, in *Proceedings of the 32nd ACM conference on hypertext and social media*, 2021, pp. 133–141.

[84] A. Lovari and S. A. Bowen, “Social media in disaster communication: A case study of strategies, barriers, and ethical implications”, *Journal of Public Affairs*, vol. 20, no. 1, e1967, 2020.

[85] Y.-J. Lu and C.-T. Li, "CGAN: Graph-aware co-attention networks for explainable fake news detection on social media", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistic,* 2020, pp. 505-514.

[86] S. Madichetty and S. Muthukumarasamy, "Detection of situational information from twitter during disaster using deep learning models", *Sādhanā*, vol. 45, pp. 1–13, 2020.

[87] S. Madichetty, S. Muthukumarasamy, and P. Jayadev, "Multi-modal classification of twitter data during disasters for humanitarian response", *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–15, 2021.

[88] S. Madichetty and M. Sridevi, "Disaster damage assessment from the tweets using the combination of statistical features and informative words", *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–11, 2019.

[89] S. Madichetty and M. Sridevi, "A neural-based approach for detecting the situational information from twitter during disaster", *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 870–880, 2021.

[90] S. Madichetty and M. Sridevi, "A novel method for identifying the damage assessment tweets during disaster", *Future Generation Computer Systems*, vol. 116, pp. 440–454, 2021.

[91] M. Martínez-Rojas, M. del Carmen Pardo-Ferreira, and J. C. Rubio-Romero, "Twitter as a tool for the management and analysis of emergency situations: A systematic literature review", *International Journal of Information Management*, vol. 43, pp. 196–208, 2018.

[92] V. V. Mahanoy, N. S. Lam, L. Zou, Z. Wang, and K. Wang, "Use of Twitter in disaster rescue: Lessons learned from Hurricane Harvey", *International Journal of Digital Earth*, vol. 13, no. 12, pp. 1454–1466, 2020.

[93] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[94] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing? understanding microblog credibility perceptions", in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 441–450.

[95] T. H. Nazer, F. Morstatter, H. Dani, and H. Liu, "Finding requests in social media for dis- aster relief", in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2016, pp. 1410–1413.

[96] D. Nguyen, K. A. Al Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks", in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 2017.

[97] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage assessment from social media imagery data during disasters", in *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, 2017, pp. 569–576.

[98] T. H. Nguyen and K. Rudra, "Towards an interpretable approach to classify and summarize crisis events from microblogs", in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3641–3650.

[99] X. Ning, L. Yao, B. Benatallah, Y. Zhang, Q. Z. Sheng, and S. S. Kanhere, "Source-aware crisis-relevant tweet identification and key information summarization", *ACM Transactions on Internet Technology (TOIT)*, vol. 19, no. 3, pp. 1–20, 2019.

[100] NIOSH, *Emergency response resources*, https://www.cdc.gov/niosh/topics/emres/natural.html, Accessed: 2023-01-08, 2019.

[101] J. R. Nurse, I. Agrafiotis, S. Creese, M. Goldsmith, and K. Lamberts, "Building confidence in information-trustworthiness metrics for decision support", in *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, IEEE, 2013, pp. 535–543.

[102] J. O'Donovan, B. Kang, G. Meyer, T. Hllerer, and S. Adali, "Credibility in context: An analysis of feature distributions in twitter", in *IEEE International Conference on Social Computing, SocialCom*, vol. 10, 2012.

[103] R. Ogie and N. Verstaevel, “Disaster informatics: An overview”, *Progress in Disaster Science,* 7, 100111, 2020.

[104] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, “Crisislex: A lexicon for collecting and filtering microblogged communications in crises”, in *Proceedings of the international AAAI conference on web and social media*, vol. 8, 2014, pp. 376–385.

[105] A. Olteanu, S. Vieweg, and C. Castillo, “What to expect when the unexpected happens: Social media communications across crises”, in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 2015, pp. 994–1009.

[106] R. Pandey, H. Purohit, J. Chan, and A. Johri, “AI for trustworthiness! credible user identification on social web for disaster response agencies”, in *AAAI FSS-18: Artificial Intelligence in Government and Public Sector Proceedings*, 2018.

[107] B. E. Parilla-Ferrer, P. L. Fernandez, and J. T. Ballena, “Automatic classification of disaster- related tweets”, in *Proc. International Conference on innovative engineering technologies (ICIET)*, vol. 62, 2014.

[108] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation”, in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[109] K. A. Qureshi, R. A. S. Malick, and M. Sabih, “Social media and microblogs credibility: Identification, theory driven framework, and recommendation”, *IEEE Access*, vol. 9, pp. 137 744–137 781, 2021.

[110] N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, “Fake news classification using trans- former based enhanced LSTM and BERT”, *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 98–105, 2022.

[111] M. Rajdev and K. Lee, “Fake and spam messages: Detecting misinformation during natural disasters on social media”, in *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, IEEE, vol. 1, 2015, pp. 17–20.

[112] B. Rath, X. Morales, and J. Srivastava, “SCARLET: Explainable attention based graph neural network for fake news spreader prediction”, in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, 2021, pp. 714–727.

[113] L. Resnyansky, “Social media data in the disaster context”, *Prometheus*, vol. 33, no. 2, pp. 187–212, 2015.

[114] S. R. Sahoo and B. B. Gupta, “Multiple features based approach for automatic fake news detection on social networks using deep learning”, *Applied Soft Computing*, vol. 100, p. 106 983, 2021.

[115] S. Salcedo-Sanz, J. Pérez-Aracil, G. Ascenso, J. Del Ser, D. Casillas-Pérez, C. Kadow, D. Fister, D. Barriopedro, R. García-Herrera, M. Giuliani, *et al.*, "Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: A review", *Theoretical and Applied Climatology*, vol. 155, no. 1, pp. 1–44, 2024.

[116] M. Sedensky, "'hell's breaking loose': A 911 center under siege by Harvey", *AP News, August*, vol. 30, 2017.

[117] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Measuring information credibility in social media using combination of user profile and message content dimensions", *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, p. 3537, 2020.

[118] A. Shevtsov, C. Tzagkarakis, D. Antonakaki, and S. Ioannidis, "Identification of twitter bots based on an explainable machine learning framework: The us 2020 elections case study", in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 956–967.

[119] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "Defend: Explainable fake news detection", in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 395–405.

[120] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media", *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[121] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective”, *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[122] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, “Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data”, in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 557–565.

[123] C. Silverman, S. Buttry, C. Wardle, *et al.*, *A definitive guide to verifying digital content for emergency coverage*, 2016.

[124] D. K. Singh, S. Shams, J. Kim, S.-j. Park, and S. Yang, “Fighting for information credibility: An end-to-end framework to identify fakenews during natural disasters.”, in *ISCRAM*, 2020, pp. 90–99.

[125] M. Singh, D. Bansal, and S. Sofat, “Who is who on twitter–spammer, fake or compromised account? a tool to reveal true identity in real-time”, *Cybernetics and Systems*, vol. 49, no. 1, pp. 1–25, 2018.

[126] N. Sitaula, C. K. Mohan, J. Grygiel, X. Zhou, and R. Zafarani, “Credibility-based fake news detection,” *Disinformation, Misinformation, and Fake News in Social Media*, p. 163, 2020.

[127] C. Song and H. Fujishiro, “Toward the automatic detection of rescue-request tweets: Analyzing the features of data verified by the press,” in *2019 International Conference on Information and Communication Technologies for*

Disaster Management (ICT-DM), IEEE, 2019, pp. 1–4.

[128] Statistica, *most expensive natural disasters in the United States as of June 2021*, https://www.statista.com/statistics/744015/most-expensive-natural-disasters-usa/, Accessed: 2021-06-06, 2021.

[129] J. Sun, "Research on the credibility of social media information based on user perception", *Security and communication networks*, vol. 2021, pp. 1–10, 2021.

[130] H. Tanev, V. Zavarella, and J. Steinberger, "Monitoring disaster impact: Detecting micro- events and eyewitness reports in mainstream and social media.", in *ISCRAM*, 2017.

[131] A. H. Tapia and K. Moore, "Good enough is good enough: Overcoming disaster response organizations' slow social media data adoption", *Computer supported cooperative work (CSCW)*, vol. 23, pp. 483–512, 2014.

[132] 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)*, IEEE, 2017, pp. 330–337.

[133] T. T. Tribune, *State says Harvey's death toll has reached 88*, https://www.texastribune. org/2017/10/13/harveys-death-toll-reaches-93-people/, Accessed: 2021-06-07, 2017.

[134] E. Twarog, *Hurricane ready: Coast guard adapts to the social media storm,* https: // www. usni. org/ magazines/ proceedings/ 2018 / october/ hurricane-ready- coast-guard-adapts-social-media-storm, Accessed: 2024-04-04, 2018.

[135] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets", *ACM Transactions on the Web (TWEB)*, vol. 13, no. 2, pp. 1–22, 2019.

[136] B. Wang and J. Zhuang, "Crisis information distribution on twitter: A content analysis of tweets during hurricane sandy", *Natural hazards*, vol. 89, pp. 161–181, 2017.

[137] Z. Wang, N. S. Lam, M. Sun, X. Huang, J. Shang, L. Zou, Y. Wu, and V. V. Mihunov, "A machine learning approach for detecting rescue requests from social media", *ISPRS International Journal of Geo-Information*, vol. 11, no. 11, p. 570, 2022.

[138] Wikipedia, *Street suffix*, https ://en.wikipedia.org/wiki/Street_suffix #United_States, Accessed: 2021-12-02.

[139] S. Yang, H. Chung, D. Singh, and S. Shams, "A two-step approach to detect and understand dismisinformation events occurring in social media: A case study with critical times", *Journal of Contingencies and Crisis Management*, vol. 31, no. 4, pp. 826–842, 2023.

[140] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web", in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 1048–1052.

[141] M. Yu, Q. Huang, H. Qin, C. Scheele, and C. Yang, "Deep learning for real-time social media text classification for situation awareness–using Hurricanes Sandy, Harvey, and Irma as case studies", *International Journal of Digital Earth*, vol. 12, no. 11, pp. 1230–1247, 2019.

[142] H. Yuan, J. Zheng, Q. Ye, Y. Qian, and Y. Zhang, "Improving fake news detection with domain-adversarial and graph-attention neural network", *Decision Support Systems*, vol. 151,
p. 113 633, 2021.

[143] Z. Yue, H. Zeng, Z. Kou, L. Shang, and D. Wang, "Contrastive domain adaptation for early misinformation detection: A case study on covid-19", in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2423–2433.

[144] 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", *Proceedings of the ACM on human-computer interaction*, vol. 2, no. CSCW, pp. 1–18, 2018.

[145] H. M. Zahera, I. A. Elgendy, R. Jalota, M. A. Sherif, E. Voorhees, and A. Ellis, "Fine-tuned BERT model for multi-label tweets classification.", in *TREC*, 2019,

pp. 1–7.

[146] K. Zahra, M. Imran, and F. O. Ostermann, “Automatic identification of eyewitness messages on twitter during disasters”, *Information processing & management*, vol. 57, no. 1, p. 102 107, 2020.

[147] X. Zheng, A. Sun, S. Wang, and J. Han, “Semi-supervised event-related tweet identification with dynamic keyword generation”, in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1619–1628.

[148] B. Zhou, L. Zou, A. Mostafavi, B. Lin, M. Yang, N. Gharaibeh, H. Cai, J. Abedin, and D. Mandal, “Victimfinder: Harvesting rescue requests in disaster response from social media with BERT”, *Computers, Environment and Urban Systems*, vol. 95, p. 101 824, 2022.

[149] L. Zou, D. Liao, N. S. Lam, M. A. Meyer, N. G. Gharaibeh, H. Cai, B. Zhou, and D. Li, “Social media for emergency rescue: An analysis of rescue requests on twitter during hurricane Harvey”, *International journal of disaster risk reduction*, vol. 85, p. 103 513, 2023.

[150] A. Zubiaga and H. Ji, “Tweet, but verify: Epistemic study of information verification on twitter”, *Social Network Analysis and Mining*, vol. 4, pp. 1–12, 2014.

APPENDIX A

PUBLIC DATA SETS USED FOR RELIABILITY ASSESSMENT

This dissertation explored publicly available datasets, particularly those focusing on detecting bots and spammers among social media users. Social media bot and spammer detection datasets are particularly valuable, as they provide insights into the characteristics of both legitimate and bot users on social media platforms. These insights are crucial for defining the appropriate metrics for the user reliability assessment component in the proposed reliability assessment model.

Bot detection data set 1 – To analyze the features that differentiate suspicious users (e.g., bots) from legitimate users, a data set that was published in [24] was used. This data set includes 8386 user accounts labeled as 'bot' and 'legitimate' users. The class distribution is detailed in Table 26. The data set includes a wide range of features collected from the users' meta-data, including the number of statuses posted by the user, followers, and friends counts, geocode information, verified status, and profile description, among many others.

Table 25. Class distribution in bot detection data set 1 [24]

Category	Bot detection data set 1
Bot user class	4912
Legitimate user class	3474
Total	8386

Bot detection data set 2 – This bot detection data set was proposed by Shevtsov et al. [118]. The authors proposed a supervised machine learning approach for bot accounts categorization in the U.S. 2020 elections. The authors have collected and annotated a large number of tweets. In total, the data set consists of 11836 users distributed as shown in Table 26.

Table 26. Class distribution in bot detection data set 2 [118]

Category	Bot detection data set 2
Bot user class	4569
Legitimate user class	7267
Total	11836

Disaster-related tweets credibility – Assery et al. [16] proposed an unsupervised learning approach to evaluate the credibility of disaster-related Twitter. In this study, 3 participants were hired to label a set of tweets collected during Hurricane Florence. The tweets were annotated based on a set of user-related and content-related criteria such as:

- Is the user trusted?
- Is the user verified?
- Does the profile description contain slang and swear words?
- Number of followers and friends
- Number of posts
- Length of the tweet

- Slang and swear words in the tweet

- Linguistic features (e.g., exclamation marks, question marks, etc)

- Number of retweets

In total, 1500 tweets were labeled into two categories: (1) credible tweets and (2) non-credible tweets. The annotators assigned a confidence level for each annotation on a 10-point scale.

APPENDIX B

RELIABILITY INDICATORS ANALYSIS FROM THE LITERATURE

In the literature, reliability has been analyzed at different dimensions [2], [23], [58], [109]: (1) medium level, (2) topic level, (3) source level, (4) post level, and (5) contextual level.

Medium credibility focuses on assessing the medium through which the message is disseminated (e.g., TV, radio, newspaper, social media platforms). Topic (event) credibility measures the level of trustworthiness associated with a specific topic or event referenced in several social media messages [23]. Source reliability examines the characteristics of the user posting the message, such as the number of followers, profile age, and number of friends, among others. Many studies presume that if the source is reliable, the message posted by this source is also reliable [109]. However, unlike conventional media such as newspapers, where the source is known, social media sources are mostly anonymous, making the process of assessing their reliability more challenging. Content credibility evaluates the features of the message itself, such as information quality, accuracy, and timeliness [78]. Contextual reliability evaluates the details surrounding the communicated message, including the time and location of the shared information. Source reliability or content reliability has been the primary focus of most of the reviewed studies in this dissertation.

Source assessment indicators – Reliability assessment of social media accounts has received a lot of attention. Several studies have developed automated solutions for identifying bots, spammers, and fake social media accounts. Detecting non-reliable users can help prevent the

propagation of rumors and misinformation. To assess the reliability of social media users, two types of features were utilized in the literature: basic features and composite features. Basic features are direct data derived from Twitter profiles' metadata, such as the number of followers and number of friends. Composite features are combinations of various metrics, such as the follower-followee ratio. Figure 43 presents the most used features as found in the reviewed studies.

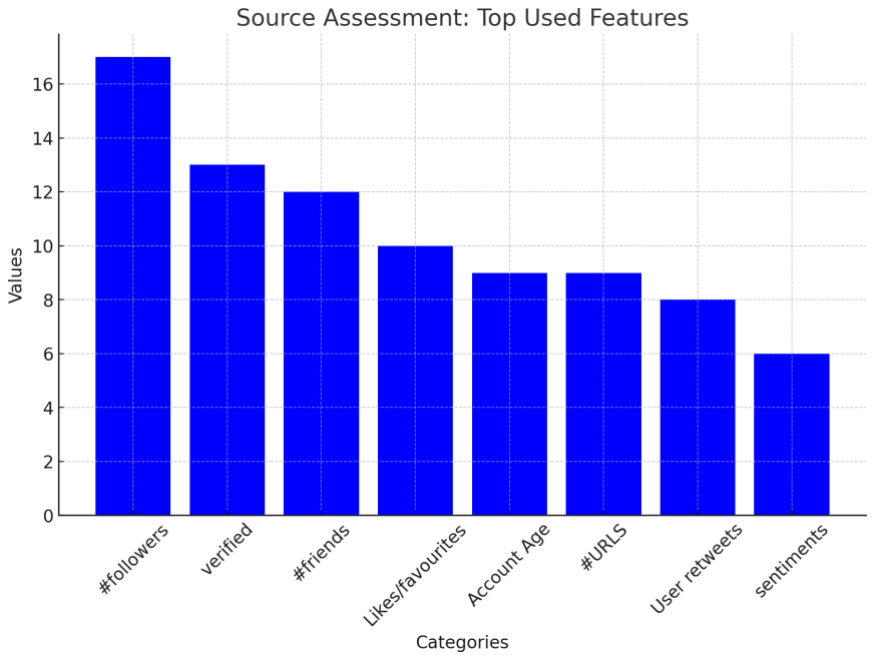

Figure 43. Top used indicators for source assessment in the literature

Other relevant features include the user's profile age, the number of URLs in the user's shared statuses, the frequency of retweets of the profile statuses, and the number of followees. Composite metrics were less commonly used in the literature. Examples of these metrics are depicted in Table 27. The identified features from the literature were categorized into several categories:

- User popularity features: this category reflects the impact that an online user has on

other users (how well-known the source is?). It can be quantified by metrics, such as the number of followers, number of friends, etc. Those features are the most common features used for social media user assessment.

- User reputation and authenticity: this category reflects the level of authenticity of a social media user. It can be quantified using metrics, such as account age and verification status. Authenticity assessment may also be derived from behavioral metrics, such as user activities, user response time, etc.
- User Expertise: This category measures the user's level of knowledge and expertise. Expertise is often topic-dependent and can be determined either from the profile's description (e.g., bio) or through analyzing the user's tweets and replies about a certain topic.

Table 27. Composite features for user assessment – examples

Feature name	metric	reference(s)
Follower-friend ratio (FFR_u)	$\frac{nb.followers - nb.friends}{Acc_{age}}$	[3]
Reputation	$\frac{nb.followers}{nb.followers + nb.followees}$	[125]
Frequency of followings	$\frac{nb.followees}{Acc_{age}}$	[125]
Follower-friend ratio	$\frac{nb.followers}{nb.friends}$	[19]

Content indicators – Content reliability measures the level of relevance and accuracy in a given social media post and is directly related to the content of the post. A well-crafted, precise message often indicates a higher level of reliability. Most previous studies have focused on implementing

machine learning and deep learning models, predominantly utilizing raw text to discern the characteristics of suspicious messages. Nonetheless, several studies have also explored text-related features within their models. Among these, linguistic features are the most used for classifying spam and fake messages. These features focus on examining the language structure of the text to detect key elements associated with suspicious messages, such as message length, the presence of question marks, and the use of pronouns, etc.

Sentiment analysis is extensively utilized to detect suspicious messages and misinformation on social media. Suspicious messages and rumors are often associated with the presence of negative words. Emotions such as fear are also good predictors of spam, rumors, and fake news. Other key content factors found in the literature include URLs, image credibility, hashtags, the number of retweets, mentions, and message location. The trustworthiness of URLs embedded in social media messages is crucial; malicious users frequently use suspicious URL links to disseminate rumors and misinformation and promote suspicious content. The presence of an image significantly enhances the reliability of an online message. However, the authenticity of the attached images should be verified. The number and types of hashtags/mentions play a critical role in evaluating the reliability of a social media post. A common behavior among spammers is to leverage mention features and hashtags to tag users and draw attention to their messages. Finally, metrics that measure the level of interaction a social media message receives, such as the number of retweets, likes, and replies, are also relevant content indicators. The most relevant content-related indicators are illustrated in Figure 44. These indicators were grouped into several classes, each of which analyzes the content of the shared message from a different angle:

- Textual features: This class focuses on analyzing the language and textual styles used

in the shared message (e.g., linguistic features)

- Engagement features: Who is interested in this information? This class focuses on analyzing the extent to which the shared message generates interaction and the popularity of the message (e.g., the number of retweets and replies).
- Sentiment features: This class focuses on conducting a sentiment analysis to analyze the emotions and sentiments in the text of the message.
- Provenance: How has the information changed? This class includes features that analyze the history of the information over time.

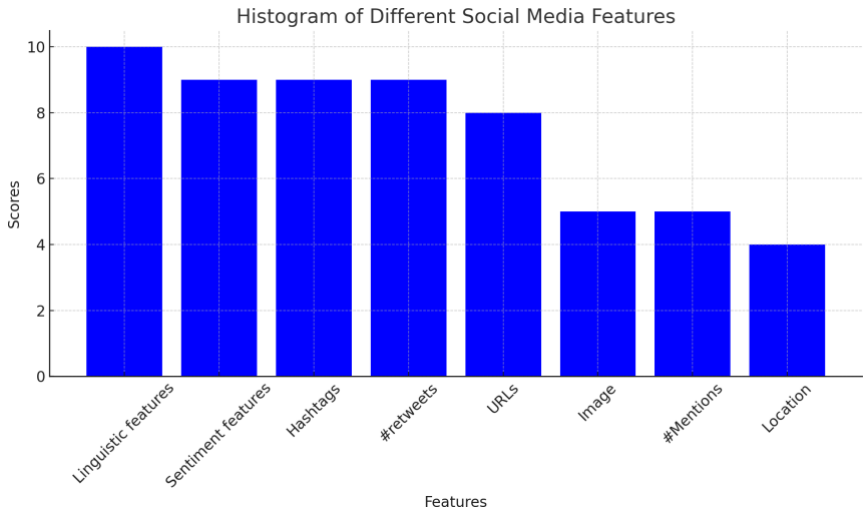

Figure 44. Top used indicators for content assessment in the literature

Contextual indicators – Contextual indicators are used to analyze the underlying context of social media information. They provide details about the time and location of the social media message and how coherent the information is with external sources. Using contextual factors to assess the credibility of online content was less common than content and source indicators. Jaho

et al. [58] proposed a credibility framework that evaluates the credibility of social media data-based on three "Cs": (1) contributor, (2) content, and (3) context. Among the contextual factors used in this framework are cross-checking, coherence, and information proximity. The posting time and location of the social media message are used as features in numerous papers. Following a discussion with a domain expert, it was highlighted that the time and location of information are important factors in determining the reliability and relevance of information on social media.

Summary – Tables 28 and 29 report the source-related features employed in previous research. Tables 30 and 31 present content-related and contextual features. Some of these indicators were integrated into the proposed reliability scoring framework.

Table 28. Source-related features from the literature

Indicator	Examples of features	reference(s)
Popularity	nb.followers	[111] [44] [68] [19] [102] [10] [124] [28] [58] [14] [39] [129] [53] [113]
	nb.followees	[102] [124] [120] [47] [85] [21] [21]
	nb.friends	[102] [28] [111] [68] [19] [101] [58] [14] [28] [114]

Table 29. Source-related features from the literature (cont.)

Indicator	Examples of features	reference(s)
Source location	geo-tagging	[111] [44] [124] [39] [12]
Authenticity	profile age	[111] [102] [10] [124] [14] [109]
	verified account	[111] [68] [19] [10] [124] [58] [14] [85] [53]
Expertise	profile description (e.g., bio)	[111] [68] [19] [10] [129] [21] [6]
	topical expertise	[94]

Table 30. Content and context-related features from the literature

Indicator	Examples of features	reference(s)
Proximity	Location description	[58] [39] [12] [6]
Linguistic features	nb.hashtags	[111] [44] [19] [102] [10] [14] [109] [114]
	URLs	[111] [44] [68] [19] [102] [124] [14] [109] [114]
Sentiments	sentimental analysis features	[44] [68] [19] [102] [46] [58] [135] [126] [109] [113]

Table 31. Content and context-related features from the literature (cont.)

Indicator	Examples of features	reference(s)
Engagement	nb.retweets	[19] [102] [10] [58] [14] [47] [85] [53] [109] [113] [94]
	nb,replies	[39] [12] [28] [124]
Time-features	publication/replying time	[111] [102] [58] [39] [12]
	recency	[44] [58]

APPENDIX C

METRICS FOR THE CREDIBILITY INDICATORS

This appendix provides details about the calculation of users' location ($GeoTag_j$) and content engagement E_i related metrics.

User's Location indicator – To determine the value to assign for the $GeoTag_j$ variable, bot detection data sets (presented in the previous appendix) were analyzed.

First, a Fisher exact test was run to analyze the degree of association between the geo-tag feature and user account type (bot/legitimate) in these data sets. The obtained odds ratio was 0.1446 (p-value = 0), indicating a strong negative association between the two variables. The likelihood of a user being geo-tagged is much lower for a bot account compared to a legitimate account. Then, the contingency table from the account type and geo-tag variables in the bot detection data sets was generated (Table 32).

Table 32. Contingency table from the bot detection data sets

	Geo-tagged	not geo-tagged
Number of bot accounts	1073	8408
Number of legitimate accounts	5035	5706

From the contingency table, the following conditional probability distributions of suspicious (e.g., bot, spam, and fake accounts) and legitimate users were calculated given the geo-tag variable.

The conditional probability distributions are shown in Table 33

Table 33. Conditional probability distributions of geo-tagged users

	P(geo-tagged=True)	P(geo-tagged=False)
P(suspicious = True | $geo-tagged$)	0.1757	0.5957
P(suspicious = False | $geo-tagged$)	0.8248	0.4043

Content engagement indicator – To assign values for the engagement metric TE^1, the posterior probability distributions of the posts' reliability were calculated by employing the Bayesian formula specified in Equation 25. To this end, Assery et al. [16] credibility data set was employed.

$$P(Credibility \mid E_{class}) = \frac{P(E_{class} \mid Credibility) * P(Credibility)}{\sum P(E_{class} \mid Credibility) * P(Credibility)} \quad (25)$$

In this equation, $P(Credibility \mid E_{class})$ denotes the posterior probability that a tweet is credible (or non-credible) given its engagement class. $P(Credibility)$ represents the likelihood of a tweet being credible (or non-credible), which is approximated from the ratio of tweets labeled as credible (or non-credible) to the total tweet count in the dataset. Lastly, $P(E_{class} \mid Credibility)$ is the prob- ability of observing a specific engagement category, given the tweet's credibility. This likelihood is calculated through the frequency of each engagement class in credible (and non-credible) and non-credible tweets. The obtained posterior probabilities calculated are shown in Table 34.

The posterior probability distributions of the posts' credibility were also calculated by employing the Bayesian formula specified in Equation 26. In this equation, $P(Credibility \mid retweet_{class})$ denotes the posterior probability that a tweet is credible (or non-credible) given its number of retweets (or retweet category). $P(Credibility)$ represents the likelihood of a tweet being credible (or non-credible), which is approximated from the ratio of tweets labeled as credible (or non-credible) to the total tweet count in the dataset. Lastly, $P(retweet_{class} \mid Credibility)$ is the probability of observing a specific retweet category, given the tweet's credibility. This likelihood is calculated through the frequency of each retweet class in credible (and non-credible) and non-credible tweets. The obtained posterior probabilities calculated are shown in Table 35.

Table 34. Posterior probabilities of posts' credibility given the engagement category

	Credible = True	**Credible = False**
$P(Credible \mid E_{C1})$	0.7035	0.2964
$P(Credible \mid E_{C2})$	0.6883	0.3116
$P(Credible \mid E_{C3})$	0.7464	0.2535
$P(Credible \mid E_{C4})$	0.5540	0.4459

$$P(Credibility \mid rt_{class}) = \frac{P(rt_{class} \mid Credibility) * P(Credibility)}{\sum P(rt_{class} \mid Credibility) * P(Credibility)} \quad (26)$$

Table 35. Posterior probabilities of posts' credibility given the retweet category

	Credible = True	**Credible = False**
$P(Credible \mid rt_{C1})$	0.6269	0.3730
$P(Credible \mid rt_{C2})$	0.7925	0.2074
$P(Credible \mid rt_{C3})$	0.7215	0.2784
$P(Credible \mid rt_{C4})$	0.8166	0.1833

VITA

EDUCATION

Ph.D. in Engineering Management and Systems Engineering **2016 – 2024**
Old Dominion University, Norfolk, VA, USA

M.S. in Computer Science **2011 – 2013**
University of Tunis, Tunisia

B.S. in Computer Science **2008 – 2011**
University of Tunis, Tunisia

RESEARCH INTERESTS

Systems Engineering, Model-based Systems Engineering, Systems Modeling and Simulation, Applied Machine Learning, and Data Analytics.

EXPERIENCE

Graduate research & Teaching Assistant **2016 - 2024**
Department of Engineering Management and Systems Engineering
Old Dominion University

System Modeling intern **2022 – 2023**
Idaho National Lab

Research Assistant **2014 – 2016**
Qatar University

PROFESSIONAL AFFILIATIONS

American Society for Engineering Management
American Society for Quality