Leveraging advanced deep learning models for disaster response

by

Soudabeh Taghian Dinani

B. S., Isfahan University of Technology, Iran, 2011

M. S., Isfahan University of Technology, Iran, 2014

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Abstract

At the beginning of natural disasters, obtaining timely and accurate information is of great importance for effective response and recovery efforts. Social media platforms like Twitter play a critical role in providing real-time updates, eyewitness accounts, and requests for assistance. Rapidly assessing this information, including identifying informative tweets and their categories, is vital for directing resources and prioritizing rescue operations. Deep learning (DL) methods have proven effective in disaster classification tasks involving both text and image data. Building upon this foundation, my research aims to further leverage advanced deep learning models to enhance the classification of disaster-related posts on X (formerly known as Twitter).

Given the challenges associated with labeling datasets in the immediate aftermath of a disaster—when data needs to be analyzed in real-time for urgent damage and situational awareness information, and manual annotation by experts is time-consuming—I have investigated methods suitable for handling smaller labeled datasets. These methods were evaluated in both in-domain and cross-domain settings to assess their effectiveness across datasets with similar and different characteristics. Additionally, the performance of the models was explored in few-shot and zero-shot settings, acknowledging the scarcity of labeled data as a disaster unfolds. In particular, the potential of Large Language Models (LLMs), due to their generalization capabilities from extensive pretraining and scale, was explored for zero-shot and few-shot settings. Results showed that LLMs are invaluable for quick response efforts in disaster situations at the beginning of a crisis.

These advancements aim to improve situational awareness and optimize resource allocation during disaster events, ultimately contributing to the reduction of human and economic tolls.

Leveraging Advanced Deep Learning Models for Disaster Response

by

Soudabeh Taghian Dinani

B.S., Isfahan University of Technology, Iran, 2011

M.S., Isfahan University of Technology, Iran, 2014

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor
Doina Caragea

Copyright

Abstract

At the beginning of natural disasters, obtaining timely and accurate information is of great importance for effective response and recovery efforts. Social media platforms like Twitter play a critical role in providing real-time updates, eyewitness accounts, and requests for assistance. Rapidly assessing this information, including identifying informative tweets and their categories, is vital for directing resources and prioritizing rescue operations. Deep learning (DL) methods have proven effective in disaster classification tasks involving both text and image data. Building upon this foundation, my research aims to further leverage advanced deep learning models to enhance the classification of disaster-related posts on X (formerly known as Twitter).

Given the challenges associated with labeling datasets in the immediate aftermath of a disaster—when data needs to be analyzed in real-time for urgent damage and situational awareness information, and manual annotation by experts is time-consuming—I have investigated methods suitable for handling smaller labeled datasets. These methods were evaluated in both in-domain and cross-domain settings to assess their effectiveness across datasets with similar and different characteristics. Additionally, the performance of the models was explored in few-shot and zero-shot settings, acknowledging the scarcity of labeled data as a disaster unfolds. In particular, the potential of Large Language Models (LLMs), due to their generalization capabilities from extensive pretraining and scale, was explored for zero-shot and few-shot settings. Results showed that LLMs are invaluable for quick response efforts in disaster situations at the beginning of a crisis.

These advancements aim to improve situational awareness and optimize resource allocation during disaster events, ultimately contributing to the reduction of human and economic tolls.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my appreciation to my advisor, Professor Doina Caragea, for her guidance and support throughout my PhD journey. Her insights and ideas have significantly contributed to the completion of this dissertation. The work presented in this dissertation would not have been possible without her patient guidance and advice.

I would also like to thank my PhD committee members, Professor Mitchell Neilsen, Professor Torben Amtoft, and Professor Caterina Scoglio, for their time, support, and for providing constructive suggestions and feedback.

I am sincerely grateful to our department head, Professor Scott DeLoach, as well as Professor Torben Amtoft and Professor Julie Thornton, for entrusting me with the role of teaching assistant for several semesters. I also thank the CS staff, ISSS staff, and Graduate School staff for their help and support. Additionally, I am grateful to my family and friends for their continuous encouragement and support.

Furthermore, I thank the National Science Foundation and Amazon Web Services for support from grant IIS-1741345, which supported part of the research and the computation in this study.

Chapter 1

Introduction

Social media has a great impact on our daily lives, and has become an important source of information in the recent years. When a disaster (such as a flood, earthquake, hurricane) strikes a place, many people use social media as a medium of communication, given that traditional emergency phone services may become unavailable due to a large volume of calls (Saroj and Pal, 2020).

Therefore, social media is of great importance in disaster response and management, and has been ranked the fourth among popular sources for acquiring emergency information (Lindsay, 2011; Kim and Hastak, 2018). This information was provided in a survey study by the American Red Cross. According to this study, the first three popular communication channels for getting information about an emergency (such as a power outage, severe weather, flash flood, hurricane, earthquake, or tornado) are TV news (with 63%), Local radio station (44%), online news (e.g., local news sites, weather.com, CNN.com) (37%), respectively; and the fourth one is Facebook (14%), a social media platform. In total, one in six (16%) have used social media to get information about an emergency[1]. Affected people in disaster zones can use social media (e.g., Twitter) to ask for help or to share emergency information (Kim and Hastak, 2018). Such information can be utilised by humanitarian organizations to facilitate post-disaster rescue and relief operations, guide the allocation of resources, and improve the real-time response overall (Alam et al., 2020a).

[1]See the American Red Cross, "Social Media in Disasters and Emergencies," blog, August 5, 2010, at http://www.slideshare.net/wharman/social-media-in-disasters-and-emergencies-aug-5.

Although such information is valuable, being posted in real-time by eyewitnesses of disasters, the amount of information posted is extremely large, and can be noisy, repetitive or even irrelevant. In fact, as soon as a disaster emerges, some people start to sympathize with the victims, express thanks to rescue organisations, repost the same content and so forth (Roy et al., 2021). Therefore, it is of great importance to effectively filter and prioritise relevant and informative content. However, manually filtering relevant information is a hard task due to the large volume and high velocity of data, and thus, automatic filtering techniques are required (Wang et al., 2021).

1.1 Deep learning for social media filtering in disasters

Deep learning has shown remarkable success across diverse application domains, ranging from natural language processing and image recognition to medical diagnosis and autonomous driving, and its effectiveness extends to disaster management, as well. These models serve as crucial automatic filtering techniques, enhancing disaster response capabilities by efficiently filtering and categorizing the vast volume of user-generated content associated with such events through social media. By leveraging the power of deep learning algorithms, we can analyze real-time data during disasters in a timely-manner, facilitating rapid identification of critical information and enhancing decision-making processes for emergency responders to improve their response operations and allocation of resources.

In this dissertation, I have explored the application of deep learning methodologies for rapidly and accurately filtering social media content, including text and images shared by affected individuals, during disaster events.

1.1.1 Crisis-related tweet classification

Recent studies have used deep learning (DL) methods for the purpose of processing crisis-related social media data for extracting relevant posts out (Roy et al., 2021; Alam et al., 2023; Khajwal et al., 2023; Sathishkumar et al., 2023; Li and Caragea, 2020; Kabir and Madria, 2019; Li et al.,

2018). In particular, the effectiveness of transformer-based models (e.g., BERT) in terms of classifying disaster tweets has been shown. For example, in a recent study by Alam et al. (2020b), BERT-type models had better performance as compared to convolutional neural networks (CNN) and FastText models (Joulin et al., 2016), and are currently considered to be state-of-the-art models for disaster tweet classification.

1.1.2 Crisis-related image classification

While early studies have focused on text analysis, more recent studies have also addressed the analysis of social media images, as images have been shown to contain useful situational awareness information that complements or adds to the information available in text (Imran et al., 2020a). Specifically, social media images provide detailed on-site information from the perspective of the eyewitnesses of the disaster (Bica et al., 2017), and can serve as an ancillary, yet rich source of visual information in disaster response. Several image and multi-modal image/text datasets have been published (Nguyen et al., 2017; Alam et al., 2018a; Mouzannar et al., 2018a), and contributed to advances in the area of image classification for disaster response. Nguyen et al. (2017) used CNNs to perform damage assessment, while Li et al. (2019) used domain adversarial neural networks (DANN) to identify damage images. Agarwal et al. (2020) proposed a multi-modal framework, called Crisis-DIAS, which uses both textual and visual information from tweets. Madichetty (2020) also used both image and text features to classify crisis-related tweets. The method proposed by Madichetty (2020) is a combination of a CNN and a standard Artificial Neural Network (ANN). The ANN was used for the text-based classification component of the model, while a fine-tuned VGG-16 architecture was used for the image-based classification component of the model. Afterwards, the late fusion technique was used to combine the output of these two models to determine if the tweet should be labeled as Informative or Non-informative. Nalluru et al. (2019) used both the semantic text and image features and combined them to classify multi-modal tweet posts. Abavisani et al. (2020) also used a multi-modal fusion approach, in order to detect crisis events for three different tasks based on the combination of information in both

images and text.

As evident from the applications discussed above, deep learning (DL) methods have proven effective in disaster classification tasks involving both text and image data. Building upon this foundation, my research aims to further leverage more advanced DL models to enhance the classification of disaster-related text and image posts on Twitter. Given the challenges associated with labeling datasets in the immediate aftermath of a disaster, which is both impractical and costly, I have investigated methods suitable for handling smaller labeled datasets. Additionally, I have explored these methods in both in-domain and cross-domain settings to assess their effectiveness across datasets with similar characteristics, from the same domain and from different domains. Furthermore, I have evaluated the performance of the models in few-shot and zero-shot settings, to account for the fact that labeled data may be scarce or entirely absent as a disaster unfolds. I will present the main contributions of my research in the next section.

1.2 Contributions of this dissertation

The contributions of this dissertations are as follows, which will be further discussed in details in Chapters 2, 3, 4, 5, respectively.

1.2.1 Disaster image classification using capsule networks

From a disaster response point of view, it is important to filter posts, in particular, text and images that provide situational awareness information, in a timely manner. For image classification, capsule networks (Sabour et al., 2017; Hinton et al., 2018) have shown superiority over convolutional neural networks (CNN). Given their success in other application domains (Kim et al., 2018; Zhao et al., 2019a; LaLonde et al., 2020; Jiménez-Sánchez et al., 2018; Afshar et al., 2020; Mobiny et al., 2020; Mittal et al., 2020), in this study, I used capsule networks to classify disaster images as Informative or Non-informative.

To complement previous related approaches that have addressed the limitations posed by the

scarcity of the labeled data, I proposed to use capsule networks for classifying disaster images posted during disasters as Informative or Non-informative. One of the main properties of capsule networks is the property of equivariance, which captures the spatial relationship between features (Patrick et al., 2019). Jiménez-Sánchez et al. (2018) used capsule networks to analyze medical images and showed that the equivariance property of capsule networks translates to models that can be effectively trained with a smaller number of labeled data points as compared to the standard convolutional neural networks. As an added benefit, they also showed that capsule networks have the ability to handle imbalanced datasets. As disaster image datasets are both small and imbalanced, capsule networks represent an attractive approach for disaster image classification. Therefore, this work presents a case study of capsule networks to disaster image classification.

Using the publicly available dataset of CrisisMMD provided by Alam et al. (2018a) (which was assembled by collecting tweets during seven natural disasters including, Hurricane Irma, Hurricane Harvey, Hurricane Maria, California Wildfires, Mexico Earthquake, Iraq-Iran Earthquake, and Sri Lanka Floods), I compared capsule network models with ResNet-18 models, for both in-domain and cross-domain settings.

The contributions for this study are as follows (and will be described in detail in Chapter 2):

- I used CapsNet models to classify disaster images. To the best of my knowledge, this study is the first to use CapsNets for disaster image classification.
- I conducted in-domain experiments and compared CapsNet models with CNN models to investigate how the CapNets handle smaller amounts of labeled data, as compared to the standard CNN models. Specifically, I used CapsNets with ResNet-18 as a backbone, and compared them with ResNet-18 baselines. The experiments were run on seven different crisis datasets.
- I also conducted cross-domain experiments using both CapsNet and ResNet-18. The purpose of these experiments was to study the ability of the CapsNet models to transfer knowl-

edge from a prior source domain to the target domain, as compared to the ResNet-18 baselines. I ran experiments on a variety of source-target disaster pairs (formed using the seven crisis datasets from the in-domain setting).

The results showed that the capsule network models had better performance for all the disaster datasets considered in the in-domain experiments, and also for most of the cross-domain pairs of disasters used in the study.

1.2.2 Disaster image classification using pre-trained transformer and contrastive learning models

In this project, I studied the use of advanced transformer and contrastive learning models for disaster image classification in a humanitarian context, with focus on state-of-the-art pre-trained vision transformers such as ViT (Dosovitskiy et al., 2020), CSWin (Dong et al., 2022) and a state-of-the-art pre-trained contrastive learning model, CLIP (Radford et al., 2021). I evaluated the performance of these models across various disaster scenarios, including in-domain and cross-domain settings, as well as few-shot learning and zero-shot learning settings. The results showed that the CLIP model outperforms the two transformer models (ViT and CSWin) and also ConvNeXts (Liu et al., 2022), a competitive CNN-based model resembling transformers, in all the settings. By improving the performance of disaster image classification, this work can contribute to the goal of reducing the number of deaths and economic losses caused by disasters, as well as helping to decrease the number of people affected by these events.

The contributions of this study are as follows (and will be described in detail in Chapter 3):

- I used pre-trained transformer models (ViT, CSWin), CLIP (the image encoder based on ViT) and ConvNeXts (a CNN-based model which resembles transformers) for disaster image classification, as shown in Figure 1.1. To the best of my knowledge, this study is the first one to use contrastive learning for disaster image classification.

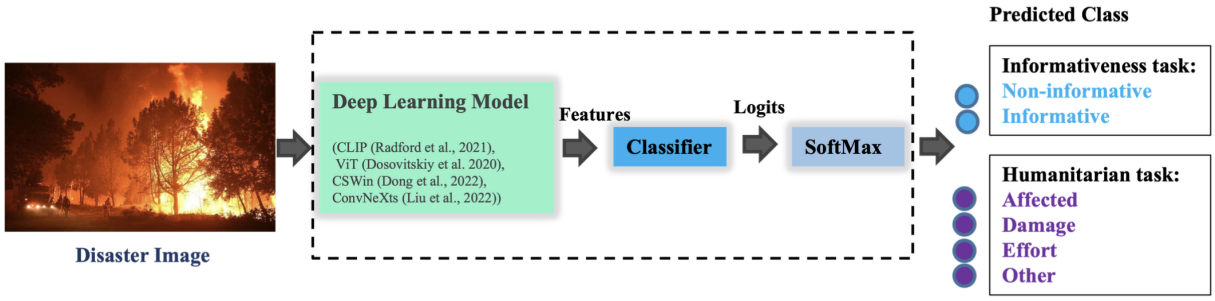


Figure 1.1: Overview of Methodology

- I studied the pre-trained CLIP model on disaster image classification in a zero-shot transfer learning setting.
- I also studied the models in a few-shot learning setting (with 1/5/10/20 instances per class, respectively) and compared the results of the few-shot models with those of the CLIP model in zero-shot setting, which can be seen as a lower-bound, as well as with the results obtained through fine-tuning the pre-trained models using all the data available, which can be seen as an upper-bound.
- I explored CLIP in several other experimental settings, including *in-domain transfer* and *cross-domain transfer*, which further involved two sub-settings of *one-versus-one* and *all-but-one*.
- The results showed that CLIP exhibits strong generalization capability across different disasters and consistently achieves the best overall performance in various settings. These findings suggest that CLIP is a strong candidate for disaster image classification tasks.

1.2.3 A comparison study for disaster tweet classification using deep learning models

Deep learning approaches, including recurrent neural networks and transformer-based models, have been previously used for disaster tweet classification (Alam et al., 2020b). Capsule Neural

Networks (CapsNets), initially proposed for image classification, have been proven to be useful for text analysis, as well. However, to the best of our knowledge, CapsNets have not been used for classifying crisis-related messages, and have not been extensively compared with state-of-the-art transformer-based models, such as BERT (Devlin et al., 2018). Therefore, in this study, I performed a thorough comparison between CapsNet models, state-of-the-art BERT models and two popular recurrent neural network models that have been successfully used for tweet classification, specifically, LSTM and Bi-LSTM models, on the task of classifying crisis tweets both in terms of their informativeness (binary classification), as well as their humanitarian content (multi-class classification). For this purpose, I used three datasets for crisis tweet classification, namely CrisisNLP (Imran et al., 2016a), CrisisLex (Olteanu et al., 2014), and CrisisBench (Alam et al., 2020b).

While CrisisNLP dataset includes tweets from 19 disasters that occurred between years 2013 and 2015 (Imran et al., 2016a), CrisisLex (Olteanu et al., 2014) is a combination of two datasets, specifically, CrisisLexT26 (consisting of tweets from 26 disasters which happened during 2012 and 2013) and CrisisLexT6 (consisting of tweets from 6 disasters which happened between October 2012 and July 2013). On the other hand, CrisisBench is a benchmark dataset constructed by Alam et al. (2020b), which is a consolidation of eight prior datasets, specifically, CrisisLex, CrisisNLP, SWDM2013, ISCRAM2013, Disaster Response Data (DRD), Disasters on Social Media (DSM), CrisisMMD, and AIDR.

The main contributions for this study are summarized as follows (and will be described in detail in Chapter 4):

- I compared CapsNets with state-of-the-art BERT-type models and also LSTM/Bi-LSTM models for text classification, including binary and multi-class classification tasks.

- I focused on the tasks of classifying crisis-related tweets in terms of both their informativeness with respect to disaster response (binary classification) and their humanitarian content (multi-class classification).

- I performed an extensive set of experiments using three datasets: CrisisLex, CrisisNLP, and CrisisBench, with the last one being a benchmark dataset consolidating eight prior datasets.

Experimental results showed that the performance of the CapsNet models is on a par with that of LSTM and Bi-LSTM models for all metrics considered, while the performance obtained with BERT models have surpassed the performance of the other three models across different datasets and classes for both classification tasks, and thus BERT could be considered the best overall model for classifying crisis tweets.

1.2.4 Exploring the potential of large language models for disaster tweet classification in zero-shot and few-shot settings

Large Language Models (LLMs) demonstrate their ability to handle complex tasks across various domains, highlighting their advanced reasoning capabilities (Nguyen et al., 2023). The study conducted by Brown et al. (2020) demonstrated that scaling up language models can lead to comparable or superior performance compared to previously leading fine-tuned models. For example, they highlighted that for NLP tasks, GPT-3 performs well in zero-shot (predicting class labels for text instances from unfamiliar categories without the need for labeled training examples) and one-shot settings and occasionally surpasses state-of-the-art fine-tuned models in few-shot scenarios (involving the model's exposure to only a few instances of labeled training examples for each class). Llama 2 (Touvron et al., 2023a) is another popular LLM, which is open-source and has been shown to be competitive with cutting-edge closed-source LLMs. The released Llama 2 family comprises pretrained and fine-tuned models, which are characterized by their extensive pretraining and expansive parameter size, yielding exceptional generalization capacities. This enables them to excel even in tasks for which they were not explicitly trained, showcasing the unique strengths of Llama 2 in zero- and few-shot learning scenarios.

In disaster scenarios, where there are limited instances available in the initial stages, employing LLMs allow us to assess their performance in zero-shot, one-shot, and few-shot settings.

Therefore, in this study, the focus is on evaluating two widely used LLMs, namely Chat-GPT(gpt-3.5-turbo) and Llama 2 for both informativeness and humanitarian tasks of three datasets of CrisisBench, CrisisNLP, and CrisisLex.

The main contributions for this study are summarized as follows (and will be described in detail in Chapter 5):

- I utilized ChatGPT and Llama 2 in zero-shot and few-shot learning settings.
- I conducted experiments with Llama 2 in zero-shot, one-shot, and five-shot settings, and with ChatGPT in zero-shot and one-shot settings for both tasks.
- I compared the results of these two LLMs with three selected fine-tuned models from the previous chapter, namely BERT (the best performer), CapsNet with static routing (chosen for its comparable performance to dynamic routing), and Bi-LSTM (slightly better than LSTM), by including their F1 score comparisons.

The experimental results and extended comparison showed that the fine-tuned BERT model remains the best for the tasks considered. However, ChatGPT and Llama 2, particularly ChatGPT, showed promising results in low-resource settings typical of post-disaster scenarios. ChatGPT performed well even with one-shot learning, outperforming Llama 2 in most zero-shot, one-shot, and five-shot settings, and nearly matching the CapsNet model. This makes ChatGPT a strong candidate for emerging crises where labeled data is scarce and real-time data filtering is needed.

1.3 Summary

Corresponding to the contributions listed above, this dissertation includes three peer-reviewed conference papers and a book chapter as follows:

- Soudabeh Taghian, Doina Caragea, Nikesh Gyawali, "Disaster Tweet Classification Using Fine-tuned Deep Learning Models Versus Zero and Few-shot Large Language Models," *in Communications in Computer and Information Science*, 2024. In press.

- Soudabeh Taghian, Doina Caragea, "Disaster Image Classification Using Pre-trained Transformer and Contrastive Learning Models," *in Proceedings of the 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1-11, October 2023. DOI: 10.1109/DSAA60987.2023.10302517

- Soudabeh Taghian, Doina Caragea, "A Comparison Study for Disaster Tweet Classification Using Deep Learning Models," *In Proceedings of the 12th International Conference on Data Science, Technology and Applications - DATA*, Vol. 1. SCITEPRESS-Science and Technology Publications, pp. 152-163, January 2023. URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0012129300003541

- Soudabeh Taghian, Doina Caragea, "Disaster Image Classification Using Capsule Networks," *in Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, July 2021. DOI: 10.1109/IJCNN52387.2021.9534448

Additionally, I performed experiments utilizing ChatGPT as a large language model to compare it with the Llamda 2 model. However, these experiments are not part of the above-listed publications and will only be reported within this dissertation.

Chapter 2

Disaster image classification using capsule networks

2.1 Introduction

Social media has a great impact on our daily lives, and has become an important source of information in the recent years. When a disaster (such as a tsunami, earthquake, flood, hurricane) strikes a place, many people use social media as a medium of communication, given that traditional emergency phone services may become unavailable due to a large volume of calls (Saroj and Pal, 2020). Therefore, social media is of great importance in disaster response and management, and has been ranked the fourth among popular sources for acquiring emergency information (Kim and Hastak, 2018). Affected people in disaster zones can use social media (e.g., Twitter) to ask for help or to share emergency information (Kim and Hastak, 2018). This information should be filtered in a timely manner, so that disaster planners and responders can use it to improve their response operations and allocation of resources.

There have been many studies on the use of deep learning approaches for filtering useful situational information from tweets. Some studies have focused on getting information from either texts (Caragea et al., 2016; Nguyen et al., 2016; Alam et al., 2018b; Taghian and Caragea, 2023a)

or images (Nguyen et al., 2017; Li et al., 2019; Imran et al., 2020b), while others have considered both text and images together (Abavisani et al., 2020; Ofli et al., 2020). For training a supervised deep learning model, a large number of data points is needed. However, in the case of a disaster, manually labeling data can be expensive and time consuming, and thus the number of data points is limited (especially in the early hours of a disaster). This problem can be somewhat alleviated by using data augmentation techniques and pre-trained models (Alom et al., 2019), or sometimes using labeled data from previous disasters, together with domain adaptation techniques, to train classifiers for a disaster of interest (Li et al., 2019; Alam et al., 2018b). It has been shown that this approach is much more useful if the previous disaster is of different nature to the current one (Li et al., 2019).

To complement previous approaches that address the limitations posed by the scarcity of the labeled data, we propose to use capsule networks for classifying disaster images posted during disasters as Informative or Non-informative. One of the main properties of capsule networks is the property of equivariance, which captures the spatial relationship between features (Patrick et al., 2019). Jiménez-Sánchez et al. (2018) used capsule networks to analyze medical images and showed that the equivariance property of capsule networks translates to models that can be effectively trained with a smaller number of labeled data points as compared to the standard convolutional neural networks. As an added benefit, they also showed that capsule networks have the ability to handle imbalanced datasets. As disaster image datasets are both small and imbalanced, capsule networks represent an attractive approach for disaster image classification. Therefore, this work presents a case study of capsule networks to disaster image classification.

Given this context, our key contributions are as follows:

- We used CapsNet models to classify disaster images. To the best of our knowledge, our study is the first to use CapsNets for disaster image classification.

- We conducted in-domain experiments and compared CapsNet models with CNN models to investigate how the CapNets handle smaller amounts of labeled data, as compared to the standard CNN models. Specifically, we used CapsNets with ResNet-18 as a backbone, and

compared them with ResNet-18 baselines. The experiments were run on seven different crisis datasets.

- We also conducted cross-domain experiments using both CapsNet and ResNet-18. The purpose of these experiments was to study the ability of the CapsNet models to transfer knowledge from a prior source domain to the target domain, as compared to the ResNet-18 baselines. We ran experiments on a variety of source-target disaster pairs (formed using the seven crisis datasets from the in-domain setting).

The rest of the chapter is organized as follows: We describe related work in Section 2.2, and provide background and approaches, including a brief description of ResNet-18 architecture and Capsule Networks, in Section 2.3. We introduce the experimental setup in Section 2.4, while presenting and discussing the results in Section 2.5. Finally, we conclude the chapter and provide ideas for future work in Section 2.6.

2.2 Related work

Works related to this chapter fall into two categories, specifically works on disaster image classification and works on capsule networks, as discussed below.

2.2.1 Disaster image classification

Many studies have used social media data to identify useful information for disaster response (Saroj and Pal, 2020). Early studies have focused on text analysis. However, more recent studies have also addressed the analysis of social media images, as images have been shown to contain useful situational awareness information that complements or adds to the information available in text (Imran et al., 2020a). Specifically, social media images provide detailed on-site information from the perspective of the eyewitnesses of the disaster (Bica et al., 2017), and can serve as an ancillary, yet rich source of visual information in disaster response.

Several image and multi-modal image/text datasets have been published (Nguyen et al., 2017; Alam et al., 2018a; Mouzannar et al., 2018a), and contributed to advances in the area of image classification for disaster response. Nguyen et al. (2017) used CNNs to perform damage assessment, while Li et al. (2019) used domain adversarial neural networks (DANN) to identify damage images. Agarwal et al. (2020) proposed a multi-modal framework, called Crisis-DIAS, which uses both textual and visual information from tweets. Madichetty (2020) also used both image and text features to classify crisis-related tweets. In the method proposed in this chapter, a combination of a CNN and a standard Artificial Neural Network (ANN) was used for the text-based classification component of the model, while the fine-tuned VGG-16 architecture was used for the image-based classification component of the model. Afterwards, the late fusion technique was used to combine the output of these two models to determine if the tweet label is Informative or Non-informative. In (Nalluru et al., 2019), both the semantic text and image features are combined to classify a multi-modal tweet post. Abavisani et al. (2020) also used a multi-modal fusion approach, in order to detect crisis events for three different tasks based on the combination of information in both images and text.

As can be seen, all these works and others that are focused on image analysis in the disaster domain have used different variants of CNN models to process images. As opposed to that, in our study, we propose to use capsule networks, which have been designed to address fundamental limitations of CNNs. In particular, we focus on the ability of the CapsNets to handle smaller labeled datasets and study them in both in-domain and cross-domain settings.

2.2.2 Capsule networks

CapsNets (Sabour et al., 2017; Hinton et al., 2018) use viewpoint invariant representations to address a fundamental limitation of CNNs, specifically, the fact that CNNs do not capture spatial relationships between features. Given their superior performance as compared to CNNs, CapsNets have been recently used in many application domains. For example, CapsNets have been used in natural language processing applications, such as text classification (Kim et al., 2018),

and multi-label text classification and question answering (Zhao et al., 2019a). They have also been used successfully in computer vision. For instance, LaLonde et al. (2020) used CapsNets for biomedical image segmentation. Jiménez-Sánchez et al. (2018) showed that CapsNet can overcome the challenges in the medical image classification domain, where datasets are usually small and imbalanced. Afshar et al. (2020) proposed a CapsNet-based framework, called COVID-CAPS, to diagnose COVID-19 cases based on X-ray images. Similarly, Mobiny et al. (2020) proposed a CapsNet model, called Detail-Oriented Capsule Networks (DECAPS), to automatically classify COVID-19 patients from Computed Tomography (CT) scans, while Mittal et al. (2020) used CapsNets to detect pneumonia from Chest X-ray images. Singh et al. (2019) used a CapsNet-based approach to create a high resolution image out of a very low resolution (VLR) image, and demonstrated the feasibility of the approach for VLR digit and face recognition.

Given the good performance of CapsNets in other application domains, especially for image classification tasks, our goal is to study the usefulness of CapsNets for disaster image classification, a task for which the available datasets are relatively small and imbalanced. To the best of our knowledge, this is the first time CapsNets are used in the disaster domain.

2.3 Background and approach

CapsNets have a CNN backbone. We use ResNet-18 as the backbone of the CapsNet models in this study, and also as a baseline when evaluating the CapsNet models. Thus, in this section, we first review the ResNet-18 architecture, and then describe the CapsNet model that we used.

2.3.1 ResNet-18 architecture

The architecture of CNN has three different types of layers including convolution layers (together with non-linear ReLU activations), sub-sampling layers (a.k.a., pooling layers), and classification layers, with the first two types of layers present in the lower levels of the network, and classification layers at the top of the network. Specifically, the architecture of a CNN consists of a sequence

of convolutional layers, interspersed with max-pooling layers, followed by fully connected (FC) layers, and finally a softmax classification layer. The outputs of the convolution and max-pooling layers can be seen as feature maps. Lower level layers correspond to more general feature maps, while the upper level layers correspond to more specific feature maps. The feature maps associated with the last convolution/max-pooling layers are provided as inputs to the classification layers, and a prediction is made based on the scores produced by the final softmax layer. In the last fully connected layer, the class with the highest score obtained from a softmax layer will be chosen as the predicted class (Alom et al., 2018).

CNN architectures have gained a lot of popularity in computer vision, due to the considerable improvements that they have produced in this field. In particular, ResNet (He et al., 2016) is a popular CNN architecture, which won the ImageNet image classification competition in 2015. The advantage of the ResNet architecture over other CNN architectures is that it incorporates identity shortcuts, which help prevent the problem of performance saturation and/or degradation, generally faced when training deeper networks. By skipping some layers, this identity shortcut connection considers the layer residual mapping instead of only the original mapping (Kwan et al., 2019). We used ResNet-18 (which has 18 layers) as the backbone of the CapsNet models in this chapter. The architecture of ResNet-18 is shown in Figure 2.1. Given that the data used in our experiments has only two classes, the FC layer of the pre-trained ResNet-18 was changed from 1000 to 2 classes.

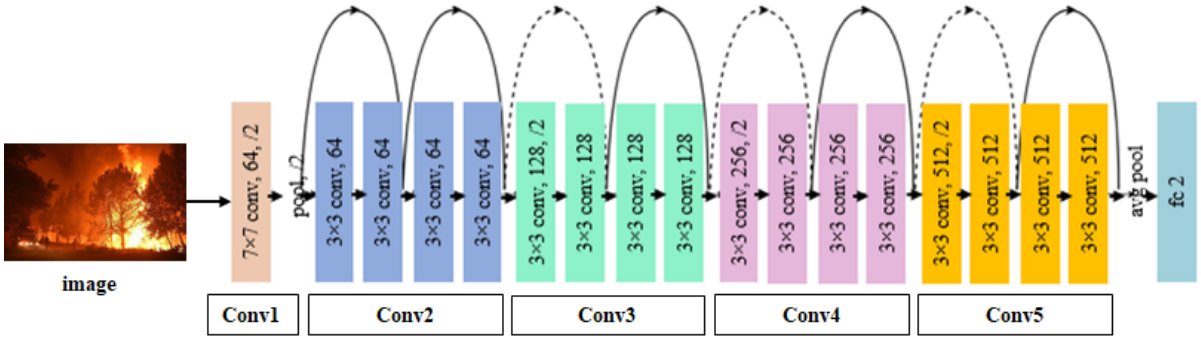


Figure 2.1: Overview of the ResNet-18 architecture

2.3.2 Capsule networks

CapsNets were first proposed by Sabour et al. (2017); Hinton et al. (2018). The main motivation for CapsNets development was given by a fundamental problem faced by standard CNNs, specifically, the fact that they are invariant to translation (due to the pooling operation), but cannot learn the spatial relationships between features. Thus, a CNN model can wrongly identify an image containing two eyes and a nose as a face, even when these features are not in the right location. Furthermore, the pooling operation results in loss of information. To compensate for this loss, CNNs need a large training dataset (Patrick et al., 2019). For example, in the case of face detection, the CNN model needs to be trained on images of faces taken from a variety of viewpoints. On the other hand, CapsNets are equivariant. For an image that shows a rotated face, a CapsNet will output the probability of the image to be classified as a face, along with the rotation degree. Therefore, CapsNets can be trained with smaller datasets and can generalize better to new unseen images with viewpoints different from those in the training set. A brief description of CapsNets is given in the following paragraph.

A CapsNet consists of several layers, with many capsules in each layer. A capsule encapsulates a group of neurons in itself. The output of each capsule can encode features of an entity (e.g., its pose, its probability of existence, its deformation, etc.) in the image. The active capsules at a lower level (the children of the higher-level capsules) vote for the capsules of the higher level (the possible parents). This is done based on transformation matrices, representing viewpoint-invariant relationship between the child capsule and the parent capsule. These transformation matrices are trained, and they change with a change in the viewpoint, making the viewpoint between the child capsule (part) and the parent capsule (whole or object) invariant. Votes are accumulated, and if several children capsules agree on a parent capsule, that capsule becomes active. This process, called routing-by-agreement, is repeated several times. The routing is done in such a way that ensures that one child capsule is routed to only one parent capsule. The result of the routing is a parsing tree with a hierarchical structure (Sabour et al., 2017; Hinton et al., 2018; Tsai et al., 2020). Dynamic routing (Sabour et al., 2017) and EM-routing (Hinton et al., 2018) are two routing

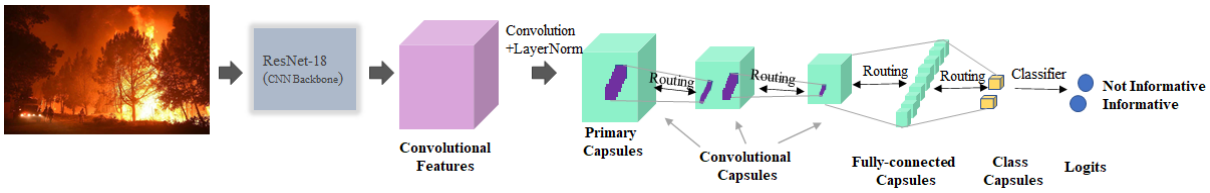


Figure 2.2: Overview of the capsule network (CapsNet) architecture, adapted from (Tsai et al., 2020)

algorithms (with some differences) that implement this mechanism.

Tsai et al. (2020) proposed a new routing algorithm, called "Inverted Dot-Product Attention Routing", which is similar to the inverted attention mechanism. In this algorithm, the parent capsules try to get attention from the children capsules. The agreement between the previous iteration vote of the parent capsule and the current iteration vote of the children capsules is determined by the routing likelihood. There are two other changes in this routing algorithm: 1) the authors used a Layer Normalization; and 2) they used concurrent iterative routing instead of sequential iterative routing; therefore, the states of the capsules and the routing likelihoods are simultaneously inferred (not in a layer-wise fashion). These changes help to scale up the model (Tsai et al., 2020).

The CapsNet architecture used in this work was adapted from (Tsai et al., 2020) and is presented in Figure 2.2. In the original architecture (Tsai et al., 2020), either several ResNet computational blocks or a single convolutional layer were used as the backbone. As opposed to that, we used ResNet-18 as our CapsNet's backbone to benefit from the pre-training of the ResNet-18 on the ImageNet dataset. For an image of size 224×224 pixels, the size of the image output by ResNet-18 is 7×7. Therefore, after the image is fed to the backbone ResNet-18, the output will be upsampled by the factor of 2 before being sent to the convolutional layer, which is followed by LayerNorm, primary capsules, two convolutional capsule layers, and finally two fully-connected capsule layers. Given that we have only two classes, the number of class capsules were changed to 2 in our model. The class probability is obtained by applying the softmax function to the resulting logits. The class with higher probability is chosen as the predicted class. As can be seen in the architecture, the routing algorithm is used from the primary capsules all the way to the class capsules (Tsai et al., 2020).

2.4 Experimental setup

In this section, the datasets used in the experiments, along with setup, evaluation metrics, and baseline are presented. The aim of the designed experiments is to answer to the following research questions:

- Does adding the capsule layers to the backbone ResNet-18 improve the performance for disaster image classification in in-domain and cross-domain settings?
- How does the performance of CapsNets on smaller datasets compare to the performance on larger datasets?
- In the cross-domain setting, how does the ability of the CapsNet to transfer information between source and target disasters compare for disasters of the same type versus disasters of different types?

To answer these questions, we selected two datasets: one comprising a diverse collection of seven natural disasters of different types—such as hurricanes, floods, fires, and earthquakes—which enabled us to simulate a variety of in-domain and cross-domain scenarios, and the other consisting of a large consolidated dataset that facilitated controlled experiments on balanced subsets of varying sizes. The details about the datasets, the experimental setup, evaluation metrics and baseline models are provided in the next subsections.

2.4.1 Datasets

CrisisMMD for general in-domain and cross-domain experiments: The dataset used in this study is called CrisisMMD (Alam et al., 2018a). This dataset was assembled by collecting tweets during seven natural disasters including, Hurricane Irma, Hurricane Harvey, Hurricane Maria, California Wildfires, Mexico Earthquake, Iraq-Iran Earthquake, and Sri Lanka Floods. The dataset includes both images and texts related to the corresponding disasters; however, we only used the images of the dataset in this study. Furthermore, while the dataset was originally designed to

Table 2.1: Data distribution of the CrisisMMD dataset for the task of identifying informative versus non-informative images.

Dataset	Event	Not-inf.	Inf.	Total	Year
D0	California Fire	604	984	1588	2017
D1	Hurricane Harvey	1982	2461	4443	2017
D2	Hurricane Irma	2303	2222	4525	2017
D3	Hurricane Maria	2330	2232	4562	2017
D4	Iraq Iran Earthquake	200	400	600	2017
D5	Mexico Earthquake	541	841	1382	2017
D6	Sri Lanka Floods	773	252	1025	2017

address several image classification tasks, we only focus on the task of classifying images in two classes: Informative and Non-informative. The detailed data distributions of the disasters in the dataset (the number of Informative and Non-informative images in each disaster) are presented in Table 2.1.

Consolidated dataset for controlled experiments on balanced subsets: The dataset used for the more controlled experiments on the balanced datasets of different sizes for the informativeness task was based on the consolidated dataset from (Alam et al., 2020a). This dataset was a consolidation of four datasets as follows:

- **Damage Assessment Dataset (DAD)** (Nguyen et al., 2017): A publicly available dataset originally labeled for damage severity levels in imagery data. Alam et al. (2020a) relabeled it for new tasks, including the informativeness task.

- **Damage Multimodal Dataset (DMD)** (Mouzannar et al., 2018b): Another publicly available dataset labeled for damage class. Alam et al. (2020a) relabeled it for new tasks, including the informativeness task.

- **AIDR Informativeness Dataset (AIDR-Info)**: Alam et al. (2020a) collected tweets and images using the AIDR system and labeled them for different tasks, including informativeness.

- **CrisisMMD** (Alam et al., 2020a): This dataset already included informativeness categories.

Table 2.2: Data splits for the consolidated dataset (Alam et al., 2020a) for informativeness task.

Class Labels / Splits	Train	Dev	Test	Total
Inf.	26486	1432	3414	31332
Not-inf.	21700	1622	5063	28385
Total	48186	3054	8477	59717

These datasets were merged to create consolidated train, development, and test sets. This large, relabeled dataset allows for controlled experiments on balanced subsets of different sizes. The detailed data distributions of this consolidated dataset are presented in Table 2.2.

2.4.2 Setup

Setup for general in-domain and cross-domain experiments: For the in-domain experiments, the dataset was randomly split into three parts: 70% as training set, 10% as the development set and 20% as the test set. Each experiment was run three times on each dataset and the average over runs were reported for each metric. For the cross-domain experiments, all the data for the source domain was used for the training, while the development set and testing set of target domain were used for hyper-parameter selection and evaluation, respectively. In order to perform a fair comparison with the in-domain results, the development and test sets are the same as those used in the in-domain experiments. Also, for the cross-domain experiments, we formed pairs of disasters based on the chronological order in which the disasters in the CrisisMMD dataset occurred. This order from the first disaster to the last one is: D6, D1, D2, D3, D5, D0, D4. Given the varying sizes of the specific disaster datasets, for some disaster pairs, the source dataset has more samples than the target dataset, while for others, the source has fewer samples, or a comparable number of samples. Furthermore, the CrisisMMD dataset consists of disasters of both similar and different types, which enables us to experiment with source and target disasters of the same type (for example, in one experiments the source data is Hurricane Harvey and the target data is Hurricane Irma), as well as disasters of different types (for example, in one experiment the source data is Sri Lanka Floods and the target data is California Fires). Thus, the diversity of pairs formed with disasters

from CrisisMMD allows us to study CapsNets under a variety of scenarios.

Each experiment, both in the in-domain and cross-domain settings, was first run on the ResNet-18 baseline (pre-trained on ImageNet). All the layers, except for Conv-5 group and FC layer, were frozen. Parameters of Conv-5 were initialized with the pre-trained values and fine-tuned for each experiment, while parameters of the FC layer were randomly initialized and trained from scratch. To identify a good set of hyper-parameters, we performed extensive tuning on the Hurricane Maria dataset, as this dataset is large compared to other datasets in CrisisMMD and it is also relatively balanced. The final hyper-parameters used in the experiments are: a learning rate of 5e-07, a weight decay (corresponding to L2 regularization) of 0.1, a batch size of 32 images, and a max number of epochs of 70. The exact number of epochs for an experiment (leading to the best model) was identified separately based on the development set in that experiment. The final performance for an experiment was evaluated using the best model on the test set corresponding to that experiment. The input images were resized to 224×224 (as the ResNet-18 model was pre-trained on images of this size).

For each experiment with the CapsNet model, the ResNet-18 model (which was tuned on the disaster dataset for the corresponding experiment in the previous part) was used as the backbone. This backbone was frozen, while the other layers of the CapsNet model were trained. The hyper-parameters selected for this part (using fine-tuning on Hurricane Maria as before) are: a learning rate of 1e-05, a weight decay of 0 (i.e., no L2 regularization), dropout with a rate of 0.3, a batch size of 32 images, and a max epoch number of 70. Similar to ResNet-18, each experiment is run for 70 epochs, but the number of epochs that result in the best accuracy on the development set is selected and used to evaluate performance on the test set. The input images for CapsNet were also resized to 224×224. The matrix format was used for the pose of each capsule, and the number of routings was 2.

The same data augmentation approach is used for the training set of all the experiments. Firstly, the image is resized to 225×225, then it is center-cropped to the size of 224×224. Then, padding of four is used and the image is randomly cropped to the size of 224×224. Afterwards, images

are flipped randomly in the horizontal direction.

Setup for controlled experiments on balanced subsets: Balanced subsets of varying sizes (1,000; 2,000; 10,000; 20,000; and 40,000 samples) were used for the training set. For the 1,000-sample training set, this comprised 500 informative and 500 non-informative images. Similarly, for the other sizes, each subset was split evenly, with half the samples being informative and half non-informative. To maintain balanced datasets, the maximum number of informative images in the development and test sets were limited to 1,432 and 3,414, respectively. Therefore, we kept the development and test sets constant at 2,800 and 6,800 samples, respectively, to ensure balance. For the ResNet-18 model, which was pre-trained on ImageNet, all layers except for the last convolutional group and fully connected (FC) layer were frozen. Parameters of the last convolutional group were fine-tuned, while the FC layer was trained from scratch. The hyperparameters were tuned separately for each training dataset size.

For the CapsNet model, the ResNet-18 (pre-trained on the disaster dataset of corresponding size) served as the backbone, with all its layers frozen. The remaining layers of the CapsNet were trained, and the hyperparameters were tuned for each dataset size.

The same data augmentation approach from previous experiments was applied to all training sets.

2.4.3 Evaluation metrics and baseline models

The metrics used are weighted precision, recall and F1-measure. We chose to focus on these metrics as some of the datasets are class imbalanced (i.e., the ratio between the number of instances in the two classes is high).

As mentioned before, the baseline used in our study is ResNet-18. Thus, the results of the CapsNet models are compared with the results of the ResNet-18 models. The disaster-tuned ResNet-18 is also used as the backbone in the CapsNet architecture, which enables us to assess the ability of the capsule layers (added after ResNet-18) to improve the performance of the baseline models.

	Resnet-18			CapsNet		
Data	Pr.	Re.	F1	Pr.	Re.	F1
D0	79.343	78.616	78.599	**83.642**	**83.543**	**83.515**
D1	77.296	77.290	77.150	**79.014**	**78.979**	**78.843**
D2	74.473	74.401	74.393	**75.256**	**75.212**	**75.188**
D3	75.731	75.731	75.724	**77.503**	**77.449**	**77.419**
D4	79.044	71.389	71.046	**82.599**	**82.778**	**82.210**
D5	76.217	75.362	75.400	**78.122**	**78.261**	**78.146**
D6	79.042	75.447	74.636	**83.850**	**84.390**	**83.735**

Table 2.3: Classification results for in-domain experiments: Precision (Pr.), Recall (Re.) and F1-measure (F1). For each metric, the best values in a row are bold-faced.

2.5 Experimental results

The results of the in-domain experiments and cross-domain experiments are shown in Tables 2.3 and 2.4, respectively. Additionally, the results of controlled experiments are presented in Table 2.6. The reported results are averaged over the three replicas of each experiment. We discuss the results with respect to our research questions, and present a brief error analysis in what follows.

2.5.1 Discussion of the results

Our first research question was focused on the ability of the capsule layers, added to the backbone ResNet-18 layers, to improve the performance on disaster image classification. As Table 2.3 shows, for all the in-domain experiments, the results of the CapsNet model are better than those of the baseline for all evaluation metrics considered. Furthermore, Table 2.4 shows that the results of the CapsNet model are also better than those of the baseline models for most cross-domain experiments. Specifically, the results obtained with CapsNet for three metrics considered (precision, recall and F1 score) are better than the ones of the ResNet-18 in 18 out of 21 cases (corresponding to different source-target combinations formed based on the chronological order). This suggests that CapsNets are capable of improving the performance on the task of classifying disaster images, as compared ResNet-18.

Our second research question was focused on the benefits of the CapsNet model as compared

		ResNet-18			CapsNet		
S	T	Pr.	Re.	F1	Pr.	Re.	F1
D1	D0	72.452	71.908	72.038	**73.894**	**73.166**	**73.324**
D2	D0	69.102	66.981	67.317	**72.450**	**69.392**	**69.818**
D3	D0	66.048	61.006	60.708	**69.981**	**67.400**	**67.850**
D5	D0	62.958	59.434	57.730	**66.049**	**66.352**	**64.541**
D6	D0	63.477	50.105	42.170	**65.344**	**50.734**	**47.36**
D6	D1	69.142	62.087	59.344	**74.001**	**68.919**	**68.291**
D1	D2	72.390	71.786	71.656	**73.789**	**73.738**	**73.736**
D6	D2	63.324	58.858	54.394	**66.038**	**62.431**	**59.836**
D1	D3	74.016	73.794	73.768	**75.241**	**75.183**	**75.176**
D2	D3	74.769	74.708	74.704	**76.368**	**76.279**	**76.276**
D6	D3	67.019	63.596	61.284	**69.405**	**66.228**	**64.499**
D0	D4	75.541	70.833	70.963	**78.995**	**76.389**	**76.923**
D1	D4	**78.809**	**77.778**	**77.860**	78.120	76.667	77.079
D2	D4	74.543	73.333	73.422	**79.142**	**75.000**	**75.652**
D3	D4	77.451	75.556	75.739	**78.460**	**75.833**	**76.412**
D5	D4	**83.714**	80.833	81.120	83.392	**82.778**	**82.954**
D6	D4	76.337	61.944	60.480	**77.717**	**63.333**	**63.527**
D1	D5	75.151	74.638	74.800	**75.405**	**75.362**	**75.349**
D2	D5	75.059	**74.155**	**74.383**	**75.623**	73.671	73.950
D3	D5	74.489	73.430	73.660	**76.268**	**73.913**	**74.150**
D6	D5	76.026	67.874	67.014	**76.623**	**68.357**	**68.124**

Table 2.4: Classification results for the cross-domain experiments: precision (Pr.), recall (Re.) and F1-measure (F1). For each metric, the best results in a row are bold-faced.

Data	Order	Size	Ratio	F1	% F1 Change
D4	1	600	2.000	**11.164**	**15.713%**
D6	2	1025	0.326	**9.099**	**12.191%**
D5	3	1382	1.555	**2.746**	**3.641%**
D0	4	1588	1.629	**4.916**	**6.254%**
D1	5	4443	1.242	1.693	2.194%
D2	6	4525	0.965	0.795	1.068%
D3	7	4562	0.958	1.695	2.238%

Table 2.5: Difference between the CapsNet and ResNet-18 mean F1-score values for in-domain experiments, and percentage change (%F1 Change). The disasters with the smallest datasets and highest observed differences are bold-faced. The Order column shows the rank of the dataset in terms of size, while the Ratio column shows the class ratio.

with the ResNet-18 baseline, when smaller or larger datasets are used for training. We answer this question in two cases: 1) in-domain experiments across CrisisMMD datasets (D0 to D6) and 2) controlled experiments on a large consolidated dataset with different sizes of training data..

Case 1: in-domain experiments across CrisisMMD datasets (D0 to D6): To facilitate the analysis of the results with respect to the dataset size, Table 2.5 shows the differences between the CapsNet and ResNet-18 mean values for the F1 metric for the in-domain experiments (with the differences for the smaller datasets shown in bold font). More specifically, each difference is calculated as the mean F1 value of CapsNet minus the corresponding F1 mean value for of ResNet-18, i.e., $F1(CapsNet) - F1(ResNet)$, according to Table 2.5. In addition to the difference, we also show the percentage change as $[F1(CapsNet) - F1(ResNet)] \div F1(ResNet)] \times 100$. The order of the datasets based on their size from the smallest to the largest is: D4, D6, D5, D0, D1, D2, and D3, as shown in Table 2.5. The table also shows the size of each dataset, and the ratio of Informative to Non-informative samples, as the class imbalance can influence the results as well. As can be seen in Table 2.5, the largest difference (with 15,713% increase in F1-score) is for D4, which is the smallest dataset (600 samples). The second largest improvement in the F1-score is observed for D6 (with 12.191% increase in F1-score), which is the next smallest dataset (1025 samples). Besides being the smallest, the two datasets also exhibit the largest class imbalance, with Informative to Non-informative ratio being 2:1 in D4, and approximately 1:3 in D6. The third and fourth largest differences are observed for D0 and D5, respectively, which are the next datasets according to size (with D5 containing 1382 samples, and D0 containing 1588 samples). While D5 is smaller than D0, D0 has a slightly higher class imbalance, and overall the improvement seen from the CapsNet is larger for D0 as compared to D5. For the larger datasets, D1, D2, D3, which have larger size (approximately, 4500 samples each) and relatively balanced class ratio, the percent increases are in the range 1-2%. Together, these results suggest that both the dataset size and the class imbalance ratio contribute to the improved performance of CapsNet. These results are consistent with the findings in prior work (Jiménez-Sánchez et al., 2018), in which it was shown that the CapsNets has a better generalization ability (compared to

CNN) in the case of small size datasets and class-imbalance. This behavior can be attributed to the equivariant property of CapsNets that can complement for limited data, and generalize better to unseen images as compared to CNN baselines.

Case 2: controlled experiments on a large consolidated dataset with different sizes of training data: Since a more controlled study, where only one of these characteristics is varied at a time, was needed to better understand the effect of each characteristic, we performed a series of controlled experiments, training both models on balanced subsets of varying sizes while keeping the development and test sets unchanged. The in-domain experiments, summarized previously, indicated that CapsNet generally outperformed ResNet-18, especially for smaller datasets and those with a higher class imbalance. This trend indicates that CapsNet's advantages are more evident in scenarios with limited data and significant class imbalance, leveraging its ability to generalize better under such conditions.

However, the controlled experiments, presented in Table 2.6, showed a different picture. When we trained both models on balanced subsets of increasing sizes, we observed that CapsNet's performance advantage over ResNet-18 was less consistent. For smaller balanced subsets (1000 and 2000 samples), ResNet-18 actually outperformed CapsNet slightly, with CapsNet showing a decrease in F1-score by up to 0.973%. This was contrary to our expectation based on the in-domain experiments. As the dataset size increased to 10,000 samples, CapsNet began to show a notable improvement, with a 2.178% increase in F1-score over ResNet-18. For even larger datasets (20,000 and 40,000 samples), CapsNet's performance improvement stabilized around 1%. These results suggest that while CapsNet can outperform ResNet-18 on larger balanced datasets, its advantage is less pronounced or even negative on smaller, balanced datasets.

Several factors could explain this discrepancy. The controlled experiments used a consolidated dataset composed of images from different kinds of disasters occurring at different times. This large, meticulously chosen dataset ensured that the training, test, and development sets were representative of a broad range of disaster scenarios. In contrast, the in-domain experiments involved datasets from specific disasters, with images from the same location, type of disaster, and time

Data Size	ResNet-18			CapsNet			%F1 Change
	Pr.	Re.	F1	Pr.	Re.	F1	
1,000	**74.599**	**73.941**	**73.766**	73.163	73.074	73.048	-0.973%
2,000	**75.071**	**74.824**	**74.761**	74.688	74.618	74.599	-0.217%
10,000	75.911	75.500	75.403	**77.208**	**77.074**	**77.045**	2.178%
20,000	78.167	77.971	77.932	**78.894**	**78.706**	**78.671**	0.948%
40,000	79.743	79.529	79.493	**80.354**	**80.338**	**80.336**	1.060%

Table 2.6: Classification results for experiments with balanced subsets of different sizes: Precision (Pr.), Recall (Re.), F1-measure (F1) and % F1 change

period. This more homogeneous data likely amplified CapsNet's ability to handle class imbalance and limited data more effectively.

The difference in dataset characteristics between the controlled and in-domain experiments highlights the importance of context in evaluating model performance. CapsNet excels in real-world scenarios where data is often limited, imbalanced, and focused on specific events, leveraging its unique architecture to achieve better generalization. However, in more controlled settings with balanced data that is representative of a broad range of disaster scenarios, this advantage diminishes. This underscores the necessity of considering dataset nature and composition in model assessments.

Our last question was related to the ability of the CapsNet model to transfer information from the source to the target, in the cross-domain setting, when the source and target disasters are of similar types or of different types, respectively. To answer this question, for each pair of disasters considered in our cross-domain experiments, Table 2.7 shows the differences between the CapsNet F1-score and the corresponding ResNet-18 F1-score (and also the percent change in F1-score). Six pairs with the largest differences and percentage change are bold-faced (the percentage change for these pairs varies from 8.399% to 15.005%). All these six cases correspond to experiments where the source and target disasters are of different types, although we want to emphasize that not all pairs with different types of disasters show a big difference. To better interpret the results, we consider cases with the same target and different sources. For the experiments with D0 (California Fires) as the target, all the sources are of types different than D0. Three of these experiments are

Source	Target	F1	%F1 Change
D1 (Hurricane)	D0 (Fires)	1.286	1.785%
D2 (Hurricane)	D0 (Fires)	2.501	3.715%
D3 (Hurricane)	D0 (Fires)	**7.142**	**11.765%**
D5 (Earthquake)	D0 (Fires)	**6.811**	**11.798%**
D6 (Floods)	D0 (Fires)	**5.190**	**12.307%**
D6 (Floods)	D1 (Hurricane)	**8.947**	**15.077%**
D1 (Hurricane)	D2 (Hurricane)	2.080	2.903%
D6 (Floods)	D2 (Hurricane)	**5.442**	**10.005**%
D1 (Hurricane)	D3 (Hurricane)	1.408	1.909%
D2 (Hurricane)	D3 (Hurricane)	1.572	2.104%
D6 (Floods)	D3 (Hurricane)	3.215	5.246%
D0 (Fires)	D4 (Earthquake)	**5.960**	**8.399%**
D1 (Hurricane)	D4 (Earthquake)	-0.781	-1.003%
D2 (Hurricane)	D4 (Earthquake)	2.230	3.037%
D3 (Hurricane)	D4 (Earthquake)	0.673	0.889%
D5 (Earthquake)	D4 (Earthquake)	1.834	2.261%
D6 (Floods)	D4 (Earthquake)	3.047	5.038%
D1 (Hurricane)	D5 (Earthquake)	0.549	0.734%
D2 (Hurricane)	D5 (Earthquake)	-0.433	-0.582%
D3 (Hurricane)	D5 (Earthquake)	0.490	0.665%
D6 (Floods)	D5 (Earthquake)	1.110	1.656%

Table 2.7: Difference between the CapsNet and ResNet-18 mean F1-score values for cross-domain experiments, and percentage change (%F1 Change). The highest observed differences are bold-faced.

among the ones with the highest F1 difference and percentage change. In particular, when the source is D3 (Hurricane Maria), D5 (Mexico Earthquake), or D6 (Sri Lanka Floods), there is an increase of 11.765%, 11.798%, and 12.307% in F1, respectively. When D2 (Hurricane Irma) is the target, the improvement in F1 is higher when D6 (Sri Lanka Floods) is the source (10.005% increase in F1) than when D1 (Hurricane Harvey) is the source (2.903% increase in F1). We observe similar results when D3 (Hurricane Maria) is the target. For the experiments with D4 (Iraq-Iran Earthquake) as the target, the two largest increases in F1 are for the cases with D0 (California Fires) and D6 (Sri Lanka Floods) as the sources. However, two other experiments with sources D3 (Hurricane Maria) and D1 (Hurricane Harvey) of types different than D4 do not have larger increase in F1 as compared to the experiment where the source is D5 (Mexico

Earthquake). Given that the size of the training dataset was observed to affect the improvements obtained with CapsNets, the results here suggest that the size of the source may also be a factor that determines the improvement, in addition to the nature of the disasters. In particular, using D5 (Mexico Earthquake) and D6 (Sri Lanka Floods) as sources generally leads to the largest increases within a group (corresponding to a target). These are the two smallest datasets in our set, supporting the hypothesis that source size is important in addition to source type. More experiments are needed to tease out the individual factors involved here.

2.5.2 Error analysis

Figure 2.3 (in-domain setting) and Figure 2.4 (cross-domain setting) show examples of images that are correctly classified by one of the models (i.e., CapsNet or ResNet-18), while they are mis-classified by the other model. The in-domain models used to classify the images in Figure 2.3 are trained and tested on California Fires (D0). Specifically, Figure 2.3 shows examples of Informative and Non-informative images correctly classified by CapsNet and mis-classified by ResNet-18 in the top row (a) and (b), and examples of Informative and Non-informative images correctly classified by ResNet and mis-classified by CapsNet in the bottom row (b) and (c). Given that the capsule network has the property of being equivariant to viewpoint, it correctly identifies image (a) as Informative with respect to California Fires, and image (b) as Non-informative. However, it seems to mistakenly classify image (d) as Informative, possibly because it identifies an orange region that resembles fire. On the other hand, it mis-classifies image (c) as Non-informative as the fire shown in that image is flattened into the image heading, while the main view of the image does not show anything related to fire. All together, the two models agree on 271 test instances (168 Informative and 103 Non-informative), and disagree on 47 instances (29 Informative and 18 Non-informative). The CapsNet model predicts correctly 33 of the images where the models disagree, while ResNet predicts correctly only 14 of those instances.

The cross-domain models used to classify the images in Figure 2.4 are trained using Sri-lanka Floods (D6) as source, and tested on Hurricane Harvey (D1) as target. CapsNet correctly

(a)

(b)

(c)

(d)

Figure 2.3: Error analysis for the in-domain experiment focused on the D0 (California Fire) dataset. (a) Example of an Informative image that is correctly classified by CapsNet, while miss-classified by ResNet-18; (b) Example of a Non-informative image that is correctly classified by CapsNet, and miss-classified by ResNet-18; (c) Example of an Informative image that is correctly classified by ResNet-18, and miss-classified by CapsNet; (d) Example of a Non-informative image correctly classified by ResNet-18, and miss-classified by CapsNet.

classifies images (a) and (b) as Informative and Non-informative, respectively, with respect to Hurricane Harvey. However, it mistakenly classifies image (c) as Non-informative, possibly because the flooded road could look as a regular road in some viewpoint, and it mis-classifies image (d) as Informative, possibly because it learns to associates big machines with recovery from a hurricane. In this experiment, the two models agree on 682 instances (334 Informative and 348 Non-Informative), and they disagree on 206 instances (158 Informative and 48 Non-informative).

(a) (b)

(c) (d)

Figure 2.4: Error analysis for the cross-domain experiment where D6 (Sri Lanka Floods) dataset is used as source and D1 (Hurricane Harvey) as target. (a) Example of an Informative image that is correctly classified by CapsNet, while miss-classified by ResNet-18; (b) Example of a Non-informative image that is correctly classified by CapsNet, and miss-classified by ResNet-18; (c) Example of an Informative image correctly classified by ResNet-18, and miss-classified by CapsNet; (d) Example of a Non-informative image correctly classified by ResNet-18 and miss-classified by CapsNet.

When in disagreement, the CapsNet model is correct in 159 cases, while the ResNet-18 model is correct in only 47 cases.

2.6 Conclusions

In this chapter, CapsNets were used with the purpose of classifying disaster images as Informative or Non-informative. Extensive in-domain and cross-domain experiments were carried out to investigate the ability of the CapsNets for this task. The results suggest that the CapsNet model has a better performance compared to the baseline ResNet-18 for all the disaster datasets for in-domain experiments. It also showed an improvement in the evaluation metrics for most of the considered cross-domain pairs of disasters. Most importantly, the results showed that the CapsNet models could improve the performance of the ResNet-18 baseline when the size of the training datasets is small, or the datasets are imbalanced. However, its relative benefit diminishes in more controlled settings, which needs more exploration and could be due to the difference in dataset characteristics between the controlled and in-domain experiments. Controlled experiments utilized a large, diverse dataset encompassing various disasters, while in-domain experiments dealt with more homogeneous datasets specific to particular types of disasters.

Chapter 3

Disaster image classification using pre-trained transformer and contrastive learning models

3.1 Introduction

Nowadays, smartphones and the internet are accessible to a majority of people and, as a result, social media platforms, such as Twitter, have become a quick and popular way of communicating and sharing information during various types of crisis situations (Ahmad et al., 2022). This is especially true during disaster events, as 911 and other emergency lines rapidly become overwhelmed by the large volume of calls from people in need of help (King, 2018). Thanks to the fast spread of news via social media, relief and response during disasters can be accelerated using this vital medium (Ahmad et al., 2022). In fact, social media can disseminate disaster-related news and updates much faster and to a broader audience as compared to traditional media. According to a report by Middle East Eye, several trapped victims were rescued during the 2023 earthquake in Turkey after posting "locations and images" on social media (Inanc, 2023). For example, a man tweeted that he and his family were buried under rubble in Hatay. Within an hour of the post, the

tweet was shared thousands of times and the family was rescued by nearby residents. Similarly, another man posted his location and a plea for help "Please help! We are under debris. There are many people here." One hour later, the man reported that "he and his mother had been rescued, but his father was not as fortunate." These examples demonstrate how social media can be used to quickly and effectively communicate information during a disaster, allowing people to share urgent needs with a wide audience and receive help (Inanc, 2023).

Information posted on social media, such as requests for help and support, damage reports or updates on relief efforts, can be utilised by humanitarian organizations to facilitate post-disaster rescue and relief operations, guide the allocation of resources, and improve the real-time response overall (Alam et al., 2020a). Although such information is valuable, being posted in real-time by eyewitnesses of disasters, the amount of information posted is extremely large, and can be noisy, repetitive or even irrelevant. In fact, as soon as a disaster emerges, some people start to sympathize with the victims, express thanks to rescue organisations, repost the same content and so forth (Roy et al., 2021). Therefore, it is of great importance to effectively filter and prioritise relevant and informative content. However, manually filtering relevant information is a hard task due to the large volume and high velocity of data, and thus, automatic filtering techniques are required (Wang et al., 2021).

For instance, Alam et al. (2023) fine-tuned convolutional neural networks (CNNs), including ResNet18, ResNet50, ResNet101, VGG16, DenseNet, SqueezeNet, MobileNet, and EfficientNet models, using disaster images for multiple tasks (disaster types, image informativeness, humanitarian categories and damage severity), and showed promising results. In another study, Banerjee et al. (2023) used GAN models to generate disaster images to augment existing datasets, and fine-tuned CNN models, such as VGG19, ResNet18, and DenseNet121, to classify different types of real-world disaster images, with ResNet18 proving to be the most effective. Furthermore, Irwansyah et al. (2023) utilized models such as PSPNet and UNet with ResNet18 and ResNet50 backbones to classify satellite disaster images into four damage classes. Collectively, these studies demonstrate the effectiveness of CNN models in disaster image classification, showcasing their

wide applicability across different image-based tasks and contexts.

However, most recently, Transformer models, such as Vision Transformer (ViT) (Dosovitskiy et al., 2020), have emerged as a powerful tool for computer vision tasks, including classification and segmentation, and have surpassed CNNs in performance (Dosovitskiy et al., 2020; Scheibenreif et al., 2022). These models have been applied to various application domains with very promising results (Han et al., 2022; Khan et al., 2022).

While research on textual analysis of crisis-related tweets has progressed significantly, research on image analysis of such tweets has not received as much attention. However, recent studies have explored the use of convolutional neural networks (CNNs) (Alam et al., 2023; Banerjee et al., 2023; Irwansyah et al., 2023) to classify disaster images into multiple categories, such as disaster types, image informativeness, humanitarian categories, and damage severity. Nevertheless, more recently, transformer models like the Vision Transformer (ViT) (Dosovitskiy et al., 2020) have demonstrated promising results, surpassing CNNs in many computer vision scenarios. For instance, a recent study (Zhao et al., 2022) used a Tokens-to-Token ViT to classify cervical cancer smear cell images. Another study (Scheibenreif et al., 2022) employed a self-supervised pre-trained Swin Transformer for classification and segmentation of land cover, and yet another study (An and Zhang, 2022) used a transformer-based model, LPViT, for defect detection and classification of printed circuit boards (PCBs).

Despite the success of transformer and contrastive learning models in various application domains, their potential for disaster image classification, especially in the case of contrastive learning, has not been much explored yet. Our study aims to investigate the feasibility of utilizing such models for this task. Given the scarcity of images in the early stages of a disaster, we plan to leverage the strengths of transformer/contrastive learning models in experimental settings such as zero-shot (Wang et al., 2019) few-shot (Wang et al., 2020) and also cross-domain settings (Xu et al., 2023a), which have been already extensively studied in the literature for other application domains (Korkmaz et al., 2022; Cheng et al., 2023; Nag et al., 2023; Xu et al., 2023b; Jiang et al., 2022; Peng et al., 2023; Huang et al., 2022; Wang et al., 2022; Sultana et al., 2022; Sun et al., 2022;

Yang et al., 2023). Specifically, we investigate the performance of transformer variants, including ViT (Dosovitskiy et al., 2020), CSWin Transformer (Dong et al., 2022), and also a contrastive learning model, the image encoder of CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021), by comparison with ConvNetXt (a modern CNN architecture designed to resemble transformers), in the context of disaster image classification.

The main contributions of this study are as follows:

- We use pre-trained transformer models (ViT, CSWin), CLIP (the image encoder based on ViT) and ConvNeXts (a CNN-based model which resembles transformers) for disaster image classification. To the best of our knowledge, our study is the first one to use contrastive learning for disaster image classification.

- We study the pre-trained CLIP model on disaster image classification in a zero-shot transfer learning setting.

- We also study the models in a few-shot learning setting (with 1/5/10/20 instances per class, respectively) and compare the results of the few-shot models with those of the CLIP model in zero-shot setting, which can be seen as a lower-bound, as well as with the results obtained through fine-tuning the pre-trained models using all the data available, which can be seen as an upper-bound.

- We explored CLIP in several other experimental settings, including *in-domain transfer* and *cross-domain transfer*, which further involved two sub-settings of *one-versus-one* and *all-but-one*.

- Our results show that CLIP exhibits strong generalization capability across different disasters and consistently achieves the best overall performance in various settings. These findings suggest that CLIP is a strong candidate for disaster image classification tasks.

Most importantly, the results of this study suggest that our research can contribute to the development of effective solutions for disaster management and response.

3.2 Related work

In this section, we review related work on CNN models used for disaster image classification as well as work on recent vision transformers and contrastive learning models.

3.2.1 Disaster image classification

While research on textual analysis of crisis-related tweets has witnessed remarkable progress, research on image analysis of such tweets has remained comparatively under-explored (Alam et al., 2023). Inspired by successes of Deep Learning (DL) in the field of image classification, some studies have been conducted on social media disaster images (Alam et al., 2020a, 2023; Li et al., 2019; Banerjee et al., 2023; Irwansyah et al., 2023; Taghian and Caragea, 2021).

For instance, Alam et al. (2023) fine-tuned convolutional neural networks (CNNs), including ResNet18, ResNet50, ResNet101, VGG16, DenseNet, SqueezeNet, MobileNet, and EfficientNet models, using disaster images for multiple tasks (disaster types, image informativeness, humanitarian categories and damage severity), and showed promising results. In another study, Banerjee et al. (2023) used GAN models to generate disaster images to augment existing datasets, and fine-tuned CNN models, such as VGG19, ResNet18, and DenseNet121, to classify different types of real-world disaster images, with ResNet18 proving to be the most effective. Furthermore, Irwansyah et al. (2023) utilized models such as PSPNet and UNet with ResNet18 and ResNet50 backbones to classify satellite disaster images into four damage classes. In (Ma et al., 2022), authors utilized average pooling to compress initial feature maps (from the convolution layer) into three distinct strips: 'row strip', 'column strip', and 'channel strip'. Applying separate attention weighting to each strip, they integrated them, resulting in the Triple-Strip Attention Mechanism (TSAM), targeted at capturing often overlooked global features by convolutions. The method was employed for both disaster image classification, distinguishing between landslide and non-landslide images, and segmentation tasks, has been shown to effectively handle landslide and flood disaster image segmentation. Collectively, these studies demonstrate the effectiveness of CNN models in disaster

image classification, showcasing their wide applicability across different image-based tasks and contexts.

However, most recently, transformer models, such as Vision Transformer (ViT) (Dosovitskiy et al., 2020), have emerged as a powerful tool for computer vision tasks, including classification and segmentation, and have surpassed CNNs in performance (Dosovitskiy et al., 2020; Scheibenreif et al., 2022). These models have been applied to various application domains with very promising results (Han et al., 2022; Khan et al., 2022; Koshy and Elango, 2023; Ma et al., 2023), as discussed in the next subsection. Another well-known contemporary model, CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021), is a contrastive learning multimodal model. Trained on a large image-text collection, CLIP demonstrates remarkable transferability, effectively accommodating diverse tasks via fine-tuning. Its distinctive zero-shot and few-shot capabilities position CLIP as a well-suited choice for disaster response settings characterized by limited initial labeled data. However, to date, CLIP has not been applied to disaster image classification, and transformer-based models remain unexplored in zero-shot, few-shot scenarios, and humanitarian task.

3.2.2 Vision transformers and contrastive learning

The introduction of vision transformers, such as ViT (Dosovitskiy et al., 2020), has brought about a paradigm shift in image analysis, particularly in the realm of image classification. Transformer-based models have established a strong presence in various image analysis domains, largely due to their superior performance compared to state-of-the-art CNNs (Dosovitskiy et al., 2020), making it increasingly difficult for CNNs to remain competitive.

In an effort to keep up with the superior performance of transformers, a "modernized" version of ResNet, known as ConvNeXts, was developed in (Liu et al., 2022). ConvNeXts was designed to resemble transformers and produced very competitive results in terms of accuracy, scalability, and robustness. At the same time, ViT models pre-trained on large datasets have resulted in outstanding performance when transferred to smaller downstream datasets (Dosovitskiy et al., 2020), and

few-shot results of ViT pre-trained on ImageNet have also been promising. For our application domain, transformer-based models pre-trained on large datasets are ideal since obtaining reliable labels requires significant time and effort, and the availability of labeled data during a disaster is limited, especially at the beginning.

Given the initial success of the ViT model, other improved versions of transformers have also emerged. For example, Swin Transformer (Liu et al., 2021), a hierarchical transformer whose representation is computed with shifted windows, was introduced and was shown to surpass the counterpart ViT on the ImageNet-1K dataset. A further developed version of Swin Transformer was presented in (Dong et al., 2022), called CSWin Transformer, in which self-attention was performed in horizontal and vertical stripes, in parallel, forming a Cross-Shaped Window self-attention. CLIP (Contrastive Language-Image Pre-training) is a popular contrastive learning multimodal model (Radford et al., 2021), which consists of an image encoder (e.g., ViT transformer architecture) and a text encoder (e.g., BERT transformer (Devlin et al., 2019)) pre-trained together on image-caption pairs using a constrastive loss that aims to produce similar representations for an image and its corresponding caption. Since CLIP was pre-trained on a large scale dataset consisting of 400 million image–text pairs (Radford et al., 2021), it has outstanding transfer capabilities and has been successfully fine-tuned to other tasks. Moreover, CLIP has also shown good zero-shot and few-shot capabilities, which makes it an ideal candidate for a disaster response setting, as labeled data is generally not available in the beginning of a disaster.

Considering the superior performance of transformer-based models in image classification tasks and their ability to transfer well to downstream tasks with datasets of small or medium size, they are a suitable option for disaster image classification tasks with limited labeled data. To this end, we study the pre-trained CLIP (Radford et al., 2021), specifically its image-encoder component, by comparison with other well-known pre-trained transformers, such as ViT (Dosovitskiy et al., 2020) and CSWin (Dong et al., 2022), for disaster image classification in a variety of settings. In addition, we use ConvNeXt (Liu et al., 2022), a CNN model that resembles hierarchical vision Transformers, as a baseline CNN.

3.3 Approaches

In this section, we briefly review the approaches used in this study, including ViT, ConvNeXts, CSWin, and CLIP.

3.3.1 Vision transformer (ViT)

ViT is a vision model based on the transformer architecture. The transformer was originally designed for natural language processing (NLP) and resulted in performance improvements for many NLP tasks (Dosovitskiy et al., 2020). The transformer utilizes a self-attention mechanism to capture long-range dependencies and to produce a global feature representation of the input (Dosovitskiy et al., 2020). In ViT, the image is split into patches of fixed-size, which are processed by a transformer encoder after being linearly embedded (Dosovitskiy et al., 2020). ViT adds a unique CLS token to the input to obtain a global representation for classification tasks (Dosovitskiy et al., 2020; Bhojanapalli et al., 2021).

3.3.2 Cross-shaped window transformer (CSWin)

CSWin is another vision model based on the transformer architecture. As opposed to ViT, which was targeted at image classification, CSWin was developed for general-purpose vision tasks (Dong et al., 2022). CSWin's main component is a Self-Attention module, which executes self-attention calculations in parallel by dividing multi-heads into parallel groups and processing the horizontal and vertical stripes simultaneously. This technique results in an efficient enlargement of the attention area for each token in a transformer block. Furthermore, by adjusting the stripe width according to the network depth, the attention area can be further increased with minimal additional computational cost. To enhance its capabilities, CSWin Transformer incorporates the Locally-enhanced Positional Encoding (LePE), which can handle different input resolutions (Dong et al., 2022).

3.3.3 ConvNeXts

ConvNeXt is a pure convolution-based model which was designed by "modernizing" the ResNet architecture to resemble a hierarchical vision transformer (Liu et al., 2022). Its architecture features depthwise convolutions and an inverted bottleneck design, which reduces Floating Point Operations (FLOPs) and enhances performance. Additionally, the incorporation of larger kernel sizes improves the network's global receptive field and accuracy. ConvNeXt has been shown to rival transformers in terms of accuracy and scalability, while maintaining the simplicity and efficiency of standard CNNs (Liu et al., 2022).

3.3.4 Contrastive language-image pre-training

CLIP is a multimodal architecture that enables joint training of NLP and computer vision encoders, by utilizing a contrastive loss. The contrastive loss encourages the model to produce similar representations for semantically related image-text pairs, while dissimilar pairs are pushed further apart in the joint image-text representation space (Radford et al., 2021). CLIP has been trained in a self-supervised manner on a dataset of 400 million image-caption pairs, with ResNet-50/ViT, respectively, as image encoders and BERT (Devlin et al., 2019) as a transformer-based text encoder. We used ViT as the image encoder in this work. Therefore, when referring to the use of the CLIP model in this chapter, we specifically mean CLIP with ViT as the image encoder (and generally refer to this model simply as CLIP). Unlike traditional approaches that rely on pre-defined object categories, CLIP learns to identify objects based on textual descriptions. This enables CLIP to excel in tasks such as zero-shot and few-shot image classification. To perform zero-shot image classification, the input image is paired one-by-one with the textual names of the categories included in the classification task (e.g., text such as "A photo of a CLASS-NAME"), and the pair is embedded with the pre-trained model. The encoded image is compared to all encoded categories to determine the category with the highest similarity (Radford et al., 2021; Khan et al., 2022).

3.4 Experimental setup

In this section, we provide an overview of the dataset used in our experiments, the experimental setup, and the evaluation metrics used to evaluate the models.

3.4.1 Dataset

We have utilized the CrisisMMD dataset (Alam et al., 2018a) for our study. This dataset consists of tweets containing images and texts from seven natural disasters, namely Hurricane Irma, Hurricane Harvey, Hurricane Maria, California Wildfires, Mexico Earthquake, Iraq-Iran Earthquake, and Sri Lanka Floods. We have focused on two different tasks available in the CrisisMMD dataset, namely Informativeness task (i.e., predicting if an image is informative with respect to a particular disaster or not informative) and Humanitarian task (i.e., predicting the humanitarian category associated with an informative image).

The Informativeness task is a binary classification task, specifically Informative (*Inf.*) and Non-informative (*Non-inf.*) with respect to a disaster of interest. The original Humanitarian task in CrisisMMD consists of eight categories, including affected-individuals, injured-or-dead-people, missing-or-found-people, infrastructure-and-utility-damage, vehicle-damage, rescue-volunteering-or-donation-effort, not-relevant-or-cant-judge, and other-relevant-information. We omitted the missing-or-found-people category due to its relatively small number of instances. We refer to the rescue-volunteering-or-donation-effort as *Effort*. Moreover, we combined the affected-individuals and injured-or-dead-people categories into one category called affected-injured-or-dead-people or simply *Affected*. Similarly, we combined infrastructure-and-utility-damage and vehicle-damage categories into one category called *Damage*. Finally, we merged the not-relevant-or-cant-judge and other-relevant-information categories into a single category called *Other*. Thus, the final dataset consists of images in four categories: *Affected*, *Damage*, *Effort* and *Other*. The distribution of datasets for the Informativeness and Humanitarian tasks are presented in Table 3.1 and 3.2, respectively.

Table 3.1: Data distribution of the CrisisMMD dataset for the informativeness task

Dataset	Event	Non-inf.	Inf.	Total
D0	California Fire	604	984	1588
D1	Hurricane Harvey	1982	2461	4443
D2	Hurricane Irma	2303	2222	4525
D3	Hurricane Maria	2330	2232	4562
D4	Iraq Iran Earthquake	200	400	600
D5	Mexico Earthquake	541	841	1382
D6	Sri Lanka Floods	773	252	1025

Table 3.2: Data distribution of the CrisisMMD dataset for the modified humanitarian task.

Dataset	Affected	Damage	Effort	Other	Total
D0	38	601	205	139	983
D1	192	1032	666	570	2460
D2	62	887	366	907	2222
D3	153	924	439	711	2227
D4	101	182	44	72	399
D5	75	210	446	104	835
D6	51	101	69	31	252

Data Splits: We randomly split each dataset into three subsets: training (70%), development (10%), and test (20%). We used three different random seeds for splitting, resulting in three different splits for the training, development, and test subsets. Each experiment was run on the three splits, and the results are reported in terms of averages over the three runs.

Note: In terms of utilizing the CrisisMMD dataset, there exist considerable variations in how researchers split the datasets, determine the number of classes for each task, and whether they present results based on the entire merged dataset or individual subsets. While Ofli et al. (2020) established standard benchmark splits for CrisisMMD tasks when utilizing the dataset for multi-modal approaches, there is an observed discrepancy in the adoption of CrisisMMD dataset across studies, particularly when focusing on a single modality (text or image) for classification purposes. Notably, when only image modality is used for classification, Alam et al. (2020a) provided benchmark splits for the combined set of all seven datasets, as opposed to individual subsets. Moreover, although humanitarian task has 8 categories, some investigations employ a modified

4-class version of CrisisMMD for experimentation, as exemplified by Alam et al. (2020a), while some others use a 5-class version, as evidenced in works such as (Ofli et al., 2020; Sirbu et al., 2022; Kotha et al., 2022). Also some works report results for all the crisis events combined (Alam et al., 2020a; Madichetty, 2020; Gautam et al., 2019), while some other studies report results for different crisis events separately. In the latter case, specific studies use all seven events (Xukun and Caragea, 2020; Madichetty et al., 2021), while others narrow their focus to a selected subset of these events (Nalluru et al., 2019; Chen et al., 2020). Thus, a comparison of prior approaches that have used the CrisisMMD dataset is hardly possible. Given this and our focus on the performance of pretrained transformers and contrastive learning CLIP model by comparison with CNNs for image classification in a variety of low-data settings, we have chosen to use classes that are well represented in the dataset (sometimes obtained by merging two of the original classes as described above). As we work with independent events, we also chose to split the data per event according to standard practices to create three random splits. This allows us to report average results over three splits (and thus capture variation in the performance of the models).

3.4.2 Evaluation metrics

In all experiments, we chose weighted precision, weighted recall, and weighted F1-score as the evaluation metrics. We present only the F1-score values in the main result tables due to limited space.

3.4.3 Experimental settings

We aim to study the effectiveness of the models considered in realistic scenarios, where little to no labeled data is generally available for an emergent disaster. Specifically, we focus on zero-shot, few-shot and in-domain/cross-domain transfer learning settings. Our baseline setting corresponds to the traditional supervised setting, where labeled data is available to train a model for a specific disaster of interest. Among the four models studied, ConvNetXt (a CNN model) is used as a baseline for the transformer-based models (ViT and CSWin) and the contrastive learning model,

CLIP. Note that we generally use the image encoder of CLIP, specifically ViT, in our experiments, although we refer to the model as CLIP. More details for the different settings included in our experimental design are provided below.

Event-specific supervised (baseline): In this setting, we train event-specific models (ConvNetXt, ViT, CSWin, and CLIP) for each of the seven datasets (events) in CrisisMMD separately. Given an event dataset D, we use the training subset of the dataset D to train (fine-tune) the pre-trained models for the Informativeness and Humanitarian tasks, respectively. We use the validation subset of D to select hyper-parameters, and the test subset of D to evaluate the performance of the model. The term "event-specific" denotes the use of a single dataset for both the training and test subsets.

Event-specific zero-shot: In the zero-shot setting, we evaluate the pre-trained CLIP model by pairing each input test image with text representing category names, one at a time (e.g., "A photo of affected, injured or dead people"). We select the category corresponding to the text with the highest similarity with the input image. The test subset of an event dataset D is used to evaluate the performance for that event. Note that we only use CLIP in the event-specific zero-shot setting as the other models do not have the ability to assign specific labels to instances without any additional fine-tuning.

Event-specific few-shot: In the few-shot setting, we use 1/5/10/20 instances per class to fine-tune models (ConvNetXt, ViT, CSWin, and CLIP) for a particular event D. The instances used are randomly selected from the training set of that event D in a cumulative fashion (by adding more instances to the previous subset). The development and test subsets of D are used to select hyper-parameters and to evaluate the performance of the models, respectively.

Based on the results of the experiments in the event-specific supervised and few-shot settings (which will be discussed in Section 3.5), we selected CLIP as a best performing model on our disaster-related tweet classification tasks and used it for additional experimentation with in-domain and cross-domain transfer settings, as described below.

In-domain transfer: In this setting, we use CLIP models trained on prior events of a specific

disaster type (called *source* events) to predict data from an emergent event of the same type (called *target* event). The fine-tuning of the models is done on the training data from the source events. The hyper-parameter selection and model evaluation are performed on the development and test subsets of the target event, respectively. There are two types of disasters with multiple events in the CrisisMMD dataset, specifically, hurricanes and earthquakes. Thus, we perform in-domain experiments using these two types of disasters independently. In each in-domain experiment, one event of a specific type is used as target, while the other event(s) are used as source.

Cross-domain transfer: In the cross-domain setting, we train models on source events of some type(s) and evaluate the models on events of potentially different type. More specifically, we consider two sub-settings here. First, we consider a **one-versus-one** setting in which the model is trained on a source event of a particular disaster type (e.g., hurricane) and evaluated on a target event of a different type (e.g., flood). This setting may be useful in situations where the source and target events happen around the same time or at the same location, or if a prior event of the same type as the target event is not available. Second, we consider the **all-but-one** setting (a.k.a., **leave-one-out** setting) where all available events (regardless of their type) are used as source to train a model for a left-out target event. This setting is helpful in understanding if more training data from a variety of events is better than a smaller amount of training data from more similar events to a specific target event. The cross-domain models are fine-tuned on the training data from the source events. The hyper-parameter selection and model evaluation are performed on the development and test subsets of the target event, respectively.

3.4.4 Research questions

Our experiments aim to answer the following research questions (RQ):

(RQ1) Among the four models considered, which model yields the best results for disaster image classification in the event-specific supervised and few-shot settings in both informativeness (binary-class) and humanitarian (multi-class) tasks?

(RQ2) How do the models perform in the few-shot setting by comparison with the supervised learning setting?

(RQ3) How does the CLIP model perform in-domain and cross-domain by comparison with few-shot/supervised settings?

(RQ4) What setting leads to the best results for an emergent disaster for which little or no labeled data is available?

(RQ5) How does the performance of these four models compare with the CapsNet model, recognized as the top-performing deep learning model in the prior chapter for the informativeness task (given the consistent utilization of identical dataset splits for train, development, and test sets across datasets of D0 to D6 in both this and the previous chapter, and considering the previous chapter's emphasis on experiments concerning the informativeness task)?

3.4.5 Hyper-parameters

A simple data augmentation approach was applied to the training set of each model. Images were resized to 225 × 225 and then randomly cropped to 224 × 224. Additionally, images were randomly flipped in the horizontal direction and normalized. For the test and development sets, images were only resized to 224 × 224 and normalized. The specific pre-trained models used in the experiments are as follows: "facebook/convnext-tiny-224" for ConvNeXts, "google/vit-base-patch16-224-in21k" for ViT, "CSWin-Tiny (CSWin_64_12211_tiny_224)" for CSWin and "ViT-B/32" for CLIP. The last layer in each pre-trained model was removed and replaced with a linear classifier (along with dropout) that had the output size equal to the number of categories in a particular task (specifically, 2 for the Informativeness task and 4 for the Humanitarian task). All layers were frozen except for the last layer (corresponding to the linear classifier), whose parameters were randomly initialized and trained from scratch. We used the Adam optimizer, a batch size of 32 and a maximum of 50 epochs for all experiments. The exact number of epochs for each experiment was determined separately based on the development set. More detail information on

Model	Settings	LR	WD	BS	DP
CLIP	In-domain	1e-4	1e-8	32	0.0
	Cross-domain	1e-4	1e-8	32	0.0
	1-Shot	8e-4	1e-7	2	0.4
	5-Shots	5e-4	1e-7	2	0.2
	10-Shots	8e-4	1e-7	4	0.4
	20-Shots	7e-4	1e-7	4	0.4
ViT	Trained by all	1e-4	0.0	32	0.0
	1-Shot	7e-4	1e-7	2	0.4
	5-Shots	7e-4	1e-7	2	0.4
	10-Shots	7e-4	1e-7	4	0.4
	20-Shots	7e-4	1e-7	4	0.4
CSWin	Trained by all	1e-4	0.0	32	0.0
	1-Shot	7e-4	1e-7	2	0.2
	5-Shots	7e-4	1e-7	2	0.3
	10-Shots	7e-4	1e-7	4	0.2
	20-Shots	7e-4	1e-7	4	0.2
ConvNeXts	Trained by all	1e-4	0.0	32	0.0
	1-Shot	8e-4	0.0	2	0.3
	5-Shots	8e-4	1e-7	2	0.2
	10-Shots	8e-4	1e-7	4	0.2
	20-Shots	8e-4	1e-7	4	0.2

Table 3.3: Hyperparameters for different models in various experimental settings for Informativeness task: learning rate (LR), weight decay (WD), training batch size (BS), and drop out (DP)

hyper-parameters including the learning rate (LR), weight decay (WD), dropout (DP), and batch size for different experiment settings can be found in Tables 3.3 and 3.4. The best model according to the development set was used to estimate the final performance on the corresponding test set.

3.5 Experimental results and discussion

In this section, we fist present the results of the experiments that we performed, followed by a thorough discussion of the results and error analysis.

Model	Settings	LR	WD	BS	DP
CLIP	In-domain	1e-4	1e-8	32	0.0
	Cross-domain	1e-4	1e-8	32	0.0
	1-Shot	8e-4	1e-7	4	0.5
	5-Shots	7e-4	1e-7	8	0.3
	10-Shots	7e-4	3e-7	4	0.3
	20-Shots	7e-4	1e-7	8	0.3
ViT	Trained by all	1e-4	0.0	32	0.0
	1-Shot	8e-4	1e-7	4	0.2
	5-Shots	8e-4	1e-7	4	0.2
	10-Shots	1e-3	1e-7	4	0.2
	20-Shots	1e-3	1e-7	8	0.1
CSWin	Trained by all	1e-4	1e-8	32	0.0
	1-Shot	7e-4	1e-7	4	0.2
	5-Shots	8e-4	1e-7	4	0.2
	10-Shots	8e-4	1e-7	4	0.3
	20-Shots	8e-4	1e-7	8	0.2
ConvNeXts	Trained by all	1e-4	1e-8	32	0.0
	1-Shot	7e-4	1e-7	4	0.3
	5-Shots	8e-4	1e-7	4	0.2
	10-Shots	1e-3	1e-7	4	0.2
	20-Shots	1e-3	1e-7	8	0.3

Table 3.4: Hyperparameters for different models in various experimental settings for Humanitarian task: learning rate (LR), weight decay (WD), training batch size (BS), and drop out (DP)

3.5.1 Results

Tables 3.5, 3.6 show the results of the comparison between the ConvNeXts, ViT, CSWin and CLIP models in event-specific supervised for both Informativeness task and Humanitarian task, respectively. Tabels 3.7, and 3.8 shows the results of the comparison between the same models in event-specific few-shot settings for both Informativeness task and Humanitarian task. The two latter tables also show the zero-shot results for the CLIP model. The tables show Precision, Recall, and F1 scores for each event and also averages over all events for the supervised setting. In the few-shot setting, the Precision, Recall, and F1 scores corresponding to k instances (shots) per class (k = 0, 1, 5, 10 and 20, respectively) are averaged over all events in the dataset. Moreover, to facilitate comparison, the F1-scores of models in event-specific supervised and few-shot settings

are shown in Figures 3.1 and 3.2, respectively. The supervised setting is used as a baseline for the few-shot setting, while the ConvNeXts is used as a baseline for the transformer/CLIP models. The best F1-score in each row is **bold-faced**.

Tables 3.9, 3.10, 3.11, 3.12, 3.13, and 3.14 present additional results for CLIP, the best performing model overall. Specifically, few shot results (with 20 instances per class), and in-domain and cross-domain transfer results are shown for each event separately, for both Informativeness and Humanitarian tasks, by comparison with the results of the supervised setting, which is the baseline setting.

3.5.2 Discussion

In what follows, we will use the results presented in in Tables 3.5, 3.6, 3.7, 3.8, 3.10, 3.11, 3.12, 3.13, and 3.14, along with the Figures 3.1 and , 3.2, to answer our research questions in Section 3.4.4.

(RQ1) *Among the four models considered, which model yields the best results for disaster image classification in the event-specific supervised and few-shot settings?* The results in Tables 3.5, 3.6, 3.7, and 3.8, along with the F1-scores of the corresponding models in Figures 3.1 and , 3.2 (to facilitate comparison), indicate that for both tasks, CLIP generally outperforms the other models in terms of F1-score by a large margin, in both supervised and few-shot settings. For example, in the supervised setting, for the informativeness task, CLIP's average F1-score over the seven events in the dataset is 86.571, while the average F1-score of the ConvNetXts baseline is 82.292. Similarly, for the humanitarian task, CLIP's average F1-score is 81.805, while ConvNetXts's average F1-score is 78.198. Similar results can be observed in the few-shot setting, where CLIP consistently and significantly outperforms ConvNetXts for both tasks. Notably, CLIP improves ConvNetXts's average F1-score by almost 10% for the informativeness task, when 5 shots are used. It is also remarkable to see that CLIP has an average F1-score of approximately 50% in the zero-shot setting for both tasks, although as expected fine-tuning the model using a small number of instances improves the performance. We believe that the superior performance of CLIP, fol-

Dataset	Metric	Models			
		CLIP	**ConvNeXts**	**CSWin**	**ViT**
D0	**Precision**	**87.018**	85.403	83.164	86.504
	Recall	**86.897**	85.430	83.229	86.373
	F1-score	**86.832**	85.335	82.962	86.269
D1	**Precision**	**87.494**	82.753	81.521	85.118
	Recall	**87.425**	82.770	81.269	85.098
	F1-score	**87.431**	82.741	81.062	85.105
D2	**Precision**	**84.895**	79.899	79.001	80.636
	Recall	**84.862**	79.816	78.784	80.626
	F1-score	**84.853**	79.808	78.763	80.622
D3	**Precision**	**84.806**	79.607	79.852	81.361
	Recall	**84.759**	79.605	79.825	81.250
	F1-score	**84.745**	79.606	79.824	81.216
D4	**Precision**	**84.669**	82.566	81.071	83.057
	Recall	**84.722**	80.833	80.833	81.667
	F1-score	**84.596**	81.236	80.915	81.896
D5	**Precision**	**86.536**	81.578	81.943	80.214
	Recall	**86.473**	81.642	82.005	80.314
	F1-score	**86.495**	81.500	81.949	80.051
D6	**Precision**	**91.038**	85.728	86.744	88.470
	Recall	**91.057**	86.179	86.179	88.780
	F1-score	**91.047**	85.819	86.392	88.412
Avg	**Precision**	**86.637**	82.505	81.899	83.623
	Recall	**86.599**	82.325	81.732	83.444
	F1-score	**86.571**	82.292	81.695	83.367

Table 3.5: Experimental results of models in supervised settings for the informativeness task. The Precision, Recall, and F1-scores for each event are reported, along with the average metrics across the seven datasets summarized in the **Avg** row. Models compared include ConvNeXts (baseline CNN model) and ViT, CSWin and CLIP (transformer-based models). The best Precision, Recall, and F1-score in each row is **bold-faced**.

lowed by ViT as the second-best model, can be attributed to several factors, including the use of the ViT model itself as the image encoder of CLIP, and also the fact that CLIP was pre-trained on a large and diverse corpus of multimodal image-caption pairs, thus enabling the model to identify more subtle image concepts that the text describes and also to generalize well to different domains and tasks.

When analyzing the results of the two transformer models (CSWin and ViT), we can see that

Dataset	Metric	Models			
		CLIP	**ConvNeXts**	**CSWin**	**ViT**
D0	**Precision**	**81.523**	80.308	78.043	74.835
	Recall	**81.973**	80.442	79.252	76.190
	F1-score	**81.047**	78.999	77.698	72.779
D1	**Precision**	**83.872**	81.418	80.149	80.970
	Recall	**83.943**	82.249	80.759	81.911
	F1-score	**82.772**	80.915	79.583	80.371
D2	**Precision**	84.440	83.612	**85.398**	84.274
	Recall	**86.336**	85.060	85.811	86.282
	F1-score	**85.097**	83.832	84.960	84.912
D3	**Precision**	**86.468**	84.163	84.190	84.882
	Recall	**86.547**	85.202	84.753	85.277
	F1-score	**86.072**	84.214	84.027	84.219
D4	**Precision**	**82.112**	79.735	78.599	80.315
	Recall	**82.083**	79.167	78.750	79.583
	F1-score	**80.531**	76.069	77.278	77.786
D5	**Precision**	**83.409**	76.351	76.455	79.348
	Recall	**82.236**	74.850	76.048	77.246
	F1-score	**80.870**	72.473	74.924	74.978
D6	**Precision**	**78.012**	76.943	66.121	76.857
	Recall	**76.667**	73.333	65.333	74.667
	F1-score	**76.243**	70.887	64.374	75.052
Avg	**Precision**	**82.834**	80.361	78.422	80.212
	Recall	**82.826**	80.043	78.672	80.165
	F1-score	**81.805**	78.198	77.549	78.585

Table 3.6: Experimental results of models in supervised settings for the humanitarian task. The Precision, Recall, and F1-scores for each event are reported, along with the average metrics across the seven datasets summarized in the **Avg** row. Models compared include ConvNeXts (baseline CNN model) and ViT, CSWin and CLIP (transformer-based models). The best Precision, Recall, and F1-score in each row is **bold-faced**.

ViT is overall better than ConvNetXts, but worse than CLIP. However, ViT's improvement over ConvNetXts is more significant for the informativeness task, while the results of the two models are somewhat comparable for the humanitarian task. Finally, CSWin seems to have the worse overall performance on the two disaster image classification tasks in our study.

(RQ2) *How do the models perform in the few-shot setting by comparison with the supervised learning setting?* Average Precision, Recall and F1 scores over the seven events in the dataset

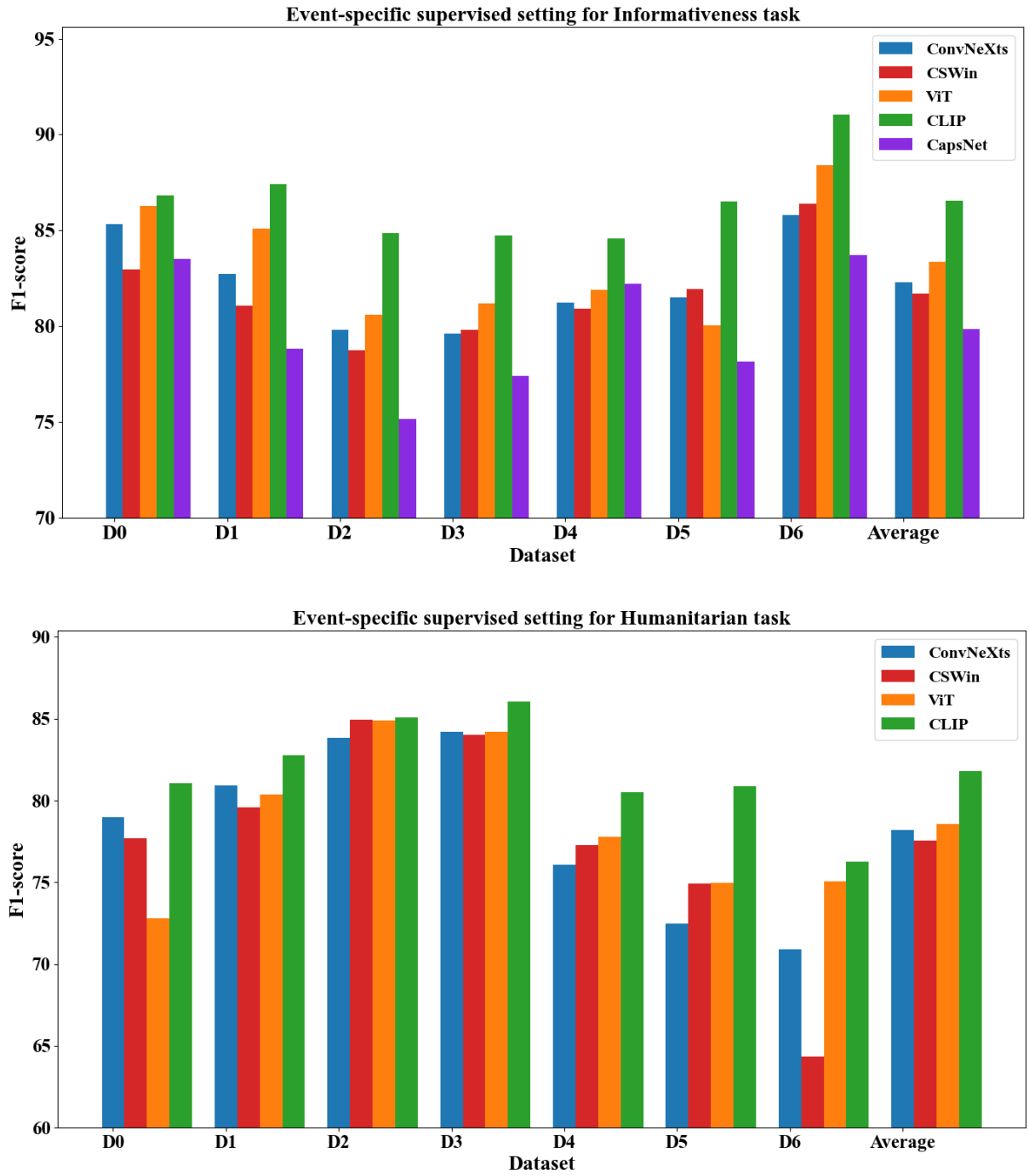


Figure 3.1: The graph for F1-score comparison of informativeness and humanitarian tasks for event-specific supervised setting

#Shots	Metric	CLIP	ConvNeXts	CSWin	ViT
0	Precision	**53.194**	-	-	-
	Recall	**50.853**	-	-	-
	F1-score	**50.816**	-	-	-
1	Precision	**69.510**	57.914	55.679	63.784
	Recall	**64.299**	55.545	49.036	61.342
	F1-score	**61.694**	55.791	50.111	61.458
5	Precision	**78.438**	69.198	69.482	71.685
	Recall	**76.548**	66.378	65.183	69.729
	F1-score	**76.418**	66.670	66.070	69.805
10	Precision	**80.383**	72.645	72.885	74.035
	Recall	**78.978**	70.097	69.535	72.340
	F1-score	**78.972**	70.292	70.252	72.416
20	Precision	**82.520**	74.899	76.480	77.597
	Recall	**81.668**	73.031	72.886	76.482
	F1-score	**81.713**	73.077	73.772	76.476
All	Precision	**86.637**	82.505	81.899	83.623
	Recall	**86.599**	82.325	81.732	83.444
	F1-score	**86.571**	82.292	81.695	83.367

Table 3.7: Experimental results of models in event-specific zero-shot and few-shot settings for the informativeness task. Precision, Recall, and F1-scores for each zero- or few-shot experiment are averaged over all events in the dataset and reported, along with the average metrics across the seven datasets from event-specific supervised settings summarized in the **All** row. Models compared include ConvNeXts (baseline CNN model) and ViT, CSWin and CLIP (transformer-based models). The best Precision, Recall, and F1-score in each row is **bold-faced**.

can be seen in Tables 3.5, and 3.6; and 3.7, 3.8, (lower part) for all models considered, in the supervised setting (All) and also in the few-shot setting (where 1, 5, 10, and 20 instances per class are used, respectively). Similarly, average F1 scores over the seven events in the dataset in the supervised learning setting can be seen in Figures 3.2, in *All (supervised)* block, for an easy visual comparison. While the results obtained in the supervised setting are better than those obtained in the few-shot setting, it is impressive to see that the results of the CLIP model fine-tuned with only 20 instances per class are relatively close to the results obtained in the supervised setting. Specifically, there is approximately 5% difference between the supervised CLIP results and the 20-shot results for both the informativeness and humanitarian tasks. It is also interesting to see that there is a significant jump from the results obtained with 1-shot versus 5-shot CLIP, especially for

#Shots	Metric	CLIP	ConvNeXts	CSWin	ViT
0	Precision	**57.791**	-	-	-
	Recall	**51.513**	-	-	-
	F1-score	**49.688**	-	-	-
1	Precision	58.261	51.989	**60.838**	56.922
	Recall	48.226	47.078	**58.417**	49.302
	F1-score	47.855	47.508	**58.285**	49.262
5	Precision	**71.200**	66.486	69.745	69.510
	Recall	**66.106**	63.475	66.797	64.299
	F1-score	**67.044**	63.792	66.746	61.694
10	Precision	**76.928**	72.280	72.601	71.967
	Recall	**73.661**	70.203	69.234	70.094
	F1-score	**74.594**	70.208	69.205	70.179
20	Precision	**79.451**	75.804	74.284	75.793
	Recall	**76.474**	73.545	72.238	74.293
	F1-score	**76.805**	73.958	72.241	73.951
All	Precision	**82.834**	80.361	78.422	80.212
	Recall	**82.826**	80.043	78.672	80.165
	F1-score	**81.805**	78.198	77.549	78.585

Table 3.8: Experimental results of models in event-specific zero-shot and few-shot settings for the humanitarian task. Precision, Recall, and F1-scores for each zero- or few-shot experiment are averaged over all events in the dataset and reported, along with the average metrics across the seven datasets from event-specific supervised settings summarized in the **All** row. Models compared include ConvNeXts (baseline CNN model) and ViT, CSWin and CLIP (transformer-based models). The best Precision, Recall, and F1-score in each row is **bold-faced**.

Dataset	Precision	Recall	F1-score
D0	83.376	83.124	82.811
D1	81.917	81.907	81.854
D2	80.100	79.705	79.650
D3	79.304	78.838	78.782
D4	82.144	80.278	80.531
D5	81.854	81.643	81.609
D6	88.944	86.179	86.751
Avg	82.520	81.668	81.713

Table 3.9: CLIP experimental results of event-specific models in event-specific few-shot settings (20 shots) for the informativeness task. Precision, Recall, and F1-scores for each event are reported, along with the average metrics across the seven datasets summarized in the **Avg** row.

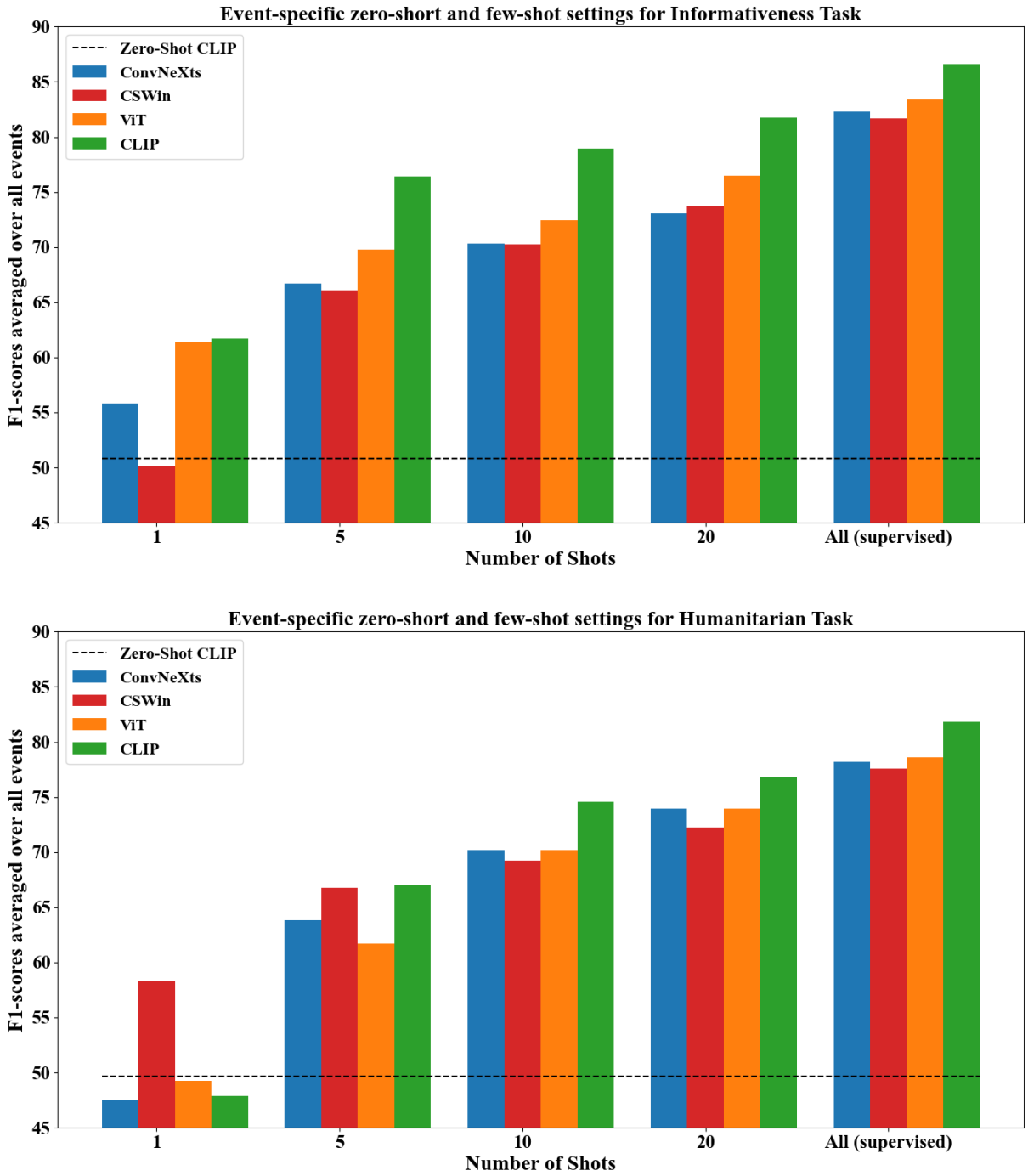


Figure 3.2: The graph for F1-score comparison of informativeness and humanitarian tasks for event-specific zero-short and few-shot settings

Dataset	Precision	Recall	F1-score
D0	82.023	77.891	78.724
D1	77.544	72.154	74.186
D2	83.842	78.979	80.737
D3	85.837	82.586	83.598
D4	81.548	81.250	81.297
D5	76.512	72.455	73.097
D6	68.848	70.000	65.997
Avg	79.451	76.474	76.805

Table 3.10: Experimental results of CLIP model in event-specific few-shot settings (20 shots) for the humanitarian task. Precision, Recall, and F1-scores for each event are reported, along with the average metrics across the seven datasets summarized in the **Avg** row.

Source	Target	Precision	Recall	F1-score
All but D0	**D0**	86.436	85.744	85.841
All but D1	**D1**	87.140	86.787	86.822
All but D2	**D2**	85.048	85.009	85.000
All but D3	**D3**	84.144	84.137	84.138
All but D4	**D4**	84.384	83.056	83.366
All but D5	**D5**	87.355	86.836	86.936
All but D6	**D6**	90.653	90.081	90.264

Table 3.11: Experimental results of CLIP model in cross-domain transfer settings (all-but-one) for the informativeness task. The Precision, Recall, and F1-scores for each event are reported in each row.

Source	Target	Precision	Recall	F1-score
All but D0	**D0**	78.286	78.061	76.415
All but D1	**D1**	81.605	82.453	81.524
All but D2	**D2**	87.227	86.712	86.294
All but D3	**D3**	84.146	84.828	83.765
All but D4	**D4**	87.354	81.667	82.81
All but D5	**D5**	74.118	65.868	65.441
All but D6	**D6**	81.996	80.000	79.034

Table 3.12: Experimental results of CLIP model in cross-domain transfer settings (all-but-one) for the humanitarian task. The Precision, Recall, and F1-scores for each event are reported in each row.

the humanitarian task. The other models also improve significantly as more instances are used and their results get close to the results of their supervised counterparts that use all the available data.

Source	Target	Precision	Recall	F1-score
D2 + D3 (Hurricane)	**D1 (Hurricane)**	86.189	85.886	85.922
D1 + D3 (Hurricane)	**D2 (Hurricane)**	85.173	85.157	85.151
D1 + D2 (Hurricane)	**D3 (Hurricane)**	84.558	84.539	84.54
D5 (Earthquake)	**D4 (Earthquake)**	86.458	85.833	85.992
D4 (Earthquake)	**D5 (Earthquake)**	82.151	82.246	82.039

Table 3.13: Experimental results of CLIP model in in-domain transfer settings (cross-event) for the informativeness task. The Precision, Recall, and F1-scores for each event are reported in each row.

Source	Target	Precision	Recall	F1-score
D2 + D3 (Hurricane)	**D1**	80.961	82.317	80.98
D1 + D3 (Hurricane)	**D2**	85.738	86.336	85.477
D1 + D2 (Hurricane)	**D3**	85.351	85.575	84.576
D5 (Earthquake)	**D4**	82.518	61.667	63.821
D4 (Earthquake)	**D5**	66.307	42.715	34.563

Table 3.14: Experimental results of CLIP model in in-domain transfer settings (cross-event) for the humanitarian task. The Precision, Recall, and F1-scores for each event are reported in each row.

Source	Target	Precision	Recall	F1-score
D3 (Hurricane)	**D0 (Fire)**	85.057	84.277	84.423
D5 (Earthquake)	**D0 (Fire)**	84.514	83.229	83.430
D6 (Flood)	**D0 (Fire)**	78.687	62.683	61.113
D0 (Fire)	**D3 (Hurricane)**	81.563	79.130	78.626
D5 (Earthquake)	**D3 (Hurricane)**	79.585	79.496	79.462
D6 (Flood)	**D3 (Hurricane)**	80.877	76.645	75.675
D0 (Fire)	**D5 (Earthquake)**	80.692	78.623	78.852
D3 (Hurricane)	**D5 (Earthquake)**	84.506	83.575	83.734
D6 (Flood)	**D5 (Earthquake)**	81.310	68.358	67.489
D0 (Fire)	**D6 (Flood)**	88.536	88.780	88.401
D3 (Hurricane)	**D6 (Flood)**	90.967	90.569	90.696
D5 (Earthquake)	**D6 (Flood)**	90.146	89.431	89.640

Table 3.15: Experimental results of CLIP model in cross-domain transfer settings (one-versus-one) for the informativeness task. The Precision, Recall, and F1-scores for each event are reported in each row.

Notably, CSWin has the best 1-shot result for the humanitarian task, but not for the informativeness task. These results together, and especially CLIP's few-shot results, show that a small number of

Source	Target	Precision	Recall	F1-score
D3 (Hurricane)	**D0 (Fire)**	78.132	77.381	74.287
D5 (Earthquake)	**D0 (Fire)**	78.038	76.701	76.283
D6 (Flood)	**D0 (Fire)**	71.920	67.347	58.872
D0 (Fire)	**D3 (Hurricane)**	77.558	81.390	78.804
D5 (Earthquake)	**D3 (Hurricane)**	82.260	80.718	80.295
D6 (Flood)	**D3 (Hurricane)**	72.652	65.022	60.798
D0 (Fire)	**D5 (Earthquake)**	66.707	58.084	56.731
D3 (Hurricane)	**D5 (Earthquake)**	72.899	62.675	61.605
D6 (Flood)	**D5 (Earthquake)**	63.367	42.515	38.179
D0 (Fire)	**D6 (Flood)**	55.520	69.333	61.457
D3 (Hurricane)	**D6 (Flood)**	79.533	78.000	75.031
D5 (Earthquake)	**D6 (Flood)**	69.529	63.333	60.317

Table 3.16: Experimental results of CLIP model in cross-domain transfer settings (one-versus-one) for the humanitarian task. The Precision, Recall, and F1-scores for each event are reported in each row.

instances from an emergent disaster event may be enough to obtain models that can be used to filter useful data in nearly real-time as a disaster emerges.

(RQ3) *How does CLIP perform in in-domain/cross-domain settings versus few-shot/supervised settings?* The results of the in-domain/cross-domain experiments and the few-shot/supervised experiments are shown in Tables 3.13, 3.14, 3.11, 3.12, 3.9, 3.10, 3.5, and 3.6 . As can be seen, the in-domain transfer and cross-domain transfer (all-but-one) results are generally better than the event-specific few-shot results but worse than the supervised results. However, there are some cases where these models produce better results than their supervised counterparts. Specifically, the in-domain transfer setting has better results than the baseline in 3 out of 14 cases, while the the cross-domain (all-but-one) setting has better results in 5 out of 14 case.

When analyzing the results of the cross-domain transfer (one-versus-one), we can see that they are sometimes better and sometimes worse than the few-shot results, depending on how similar the source and target events are in terms of disaster scene (e.g., Hurricane Maria/Sri Lanka Floods) or the time when they happened (e.g., Hurricane Maria/Mexico Earthquake), among others.

(RQ4) *What setting leads to the best results for an emergent disaster for which little or no labeled data is available?* While models trained/fine-tuned in the supervised setting give com-

petitive results overall, supervised models fine-tuned with a relatively large amount of labeled data are not practical to use for an emergent disaster as labeled data is simply not available in the early hours of the disaster. Considering the other more practical settings, the best non-supervised results for the informativeness task are split between the in-domain transfer and cross-domain transfer (all-but-one) settings, which assume no labeled data from the target disaster. However, even when the in-domain transfer results are better, the cross-domain transfer (one-but-all) models follow closely behind, making this type of model a strong candidate for filtering informative tweets in the early hours of a disaster event. Similarly, the cross-domain (all-but-one) models are also strong candidates for being used to filter tweets according to various humanitarian categories in the early hours of a disaster.

(RQ5) *How does the performance of these four models compare with the CapsNet model, recognized as the top-performing deep learning model in the prior chapter for the informativeness task (given the consistent utilization of identical dataset splits for train, development, and test sets across datasets of D0 to D6 in both this and the previous chapter, and considering the previous chapter's emphasis on experiments concerning the informativeness task)?* To address this research question, a visual comparison was conducted to avoid redundancy with the results from the previous chapter. The F1-scores of the CapsNet model across datasets D0 to D6 were integrated into the informativeness graph in Figure 3.1. As depicted in the graph, it is evident that CLIP consistently outperforms the other models, including CapsNet, across all datasets. Notably, the F1-scores of CLIP are substantially higher than that of CapsNet, particularly on datasets D2, D1, and D5.

For ConvNeXts, CSWin, and ViT, their F1-scores are mostly higher than CapsNet with two exceptions. Specifically, on dataset D0, CapsNet achieved an F1 score of 83.515, slightly surpassing CSWin with an F1 score of 82.962. Additionally, on dataset D4, CapsNet's F1 score of 82.210 was marginally higher than ConvNeXts, CSWin, and ViT, with F1 scores of 81.236, 80.915, and 81.896 respectively.

This suggests that in the early stages of a disaster, using an off-the-shelf CLIP models previ-

ously fine-tuned using as many images as available from diverse prior disasters may help achieve high performance on the target disaster. In fact, the performance may be similar to what one might obtain when using an event-specific supervised setting. As the disaster unfolds and some labeled instances from the target disaster become available, one may also consider fine-tuning a pre-trained CLIP model using directly instances from the target disaster.

3.6 Error analysis

We focus our error analysis on the humanitarian classification task. To understand what types of errors that different models make for different types of events, in Figure 3.3, we show confusion matrices for the best performing models, specifically, event-specific supervised models (left), event-specific 20-shot models (middle) and all-but-one models (right). From top to bottom, results are shown for a fire (D0), a hurricane (D3), an earthquake (D5) and a flood (D6). Based on the analysis of these confusion matrices, combined with analysis of correctly classified and misclassified images shown in Figure 3.4, several trends can be observed:

- The models often confuse the categories of *Affected* or *Effort* and *Damage*, especially when there are both humans/cars and signs of destruction in the image (for instance, images (d), (e), (m), (n), (q), (r), and (s) in Figure 3.4). This confusion arises because both categories can be inferred from the image. Particularly, if the destruction occupies a larger portion of the image, it is more likely to be categorized as *Damage* even if the label for the image is *Affected* or *Effort*. In some cases, multiple labels can be considered valid, as the image is related to damage, and there are people in the image affected by the disaster or making efforts to help the victims. However, each image in the dataset has only one label.

- Another issue identified, especially in the few-shot setting, is that the models have difficulty recognizing *Effort* images. These images are often categorized as *Affected* since there are people present in the image (for instance, image (l) in Figure 3.4). The models may need more exposure to instances where specific items such as hats, gloves, or uniforms are worn

in *Effort* images to better understand their distinctive characteristics. In contrast, it is interesting to see that the few-shot models have best performance on the *Affected* class for the fire, hurricane and earthquake events.

- In *Effort* images, dogs are commonly present (for search and rescue purposes). Consequently, models sometimes misclassify images with dogs as *Effort* even if it is evident that the dogs are in distress (for instance, image (j) in Figure 3.4).

- Additionally, the model tends to incorrectly classify images with a significant amount of red or orange color as *Damage* since it associates these colors with fire. This association leads to misclassifications when the color is present in contexts unrelated to fire.

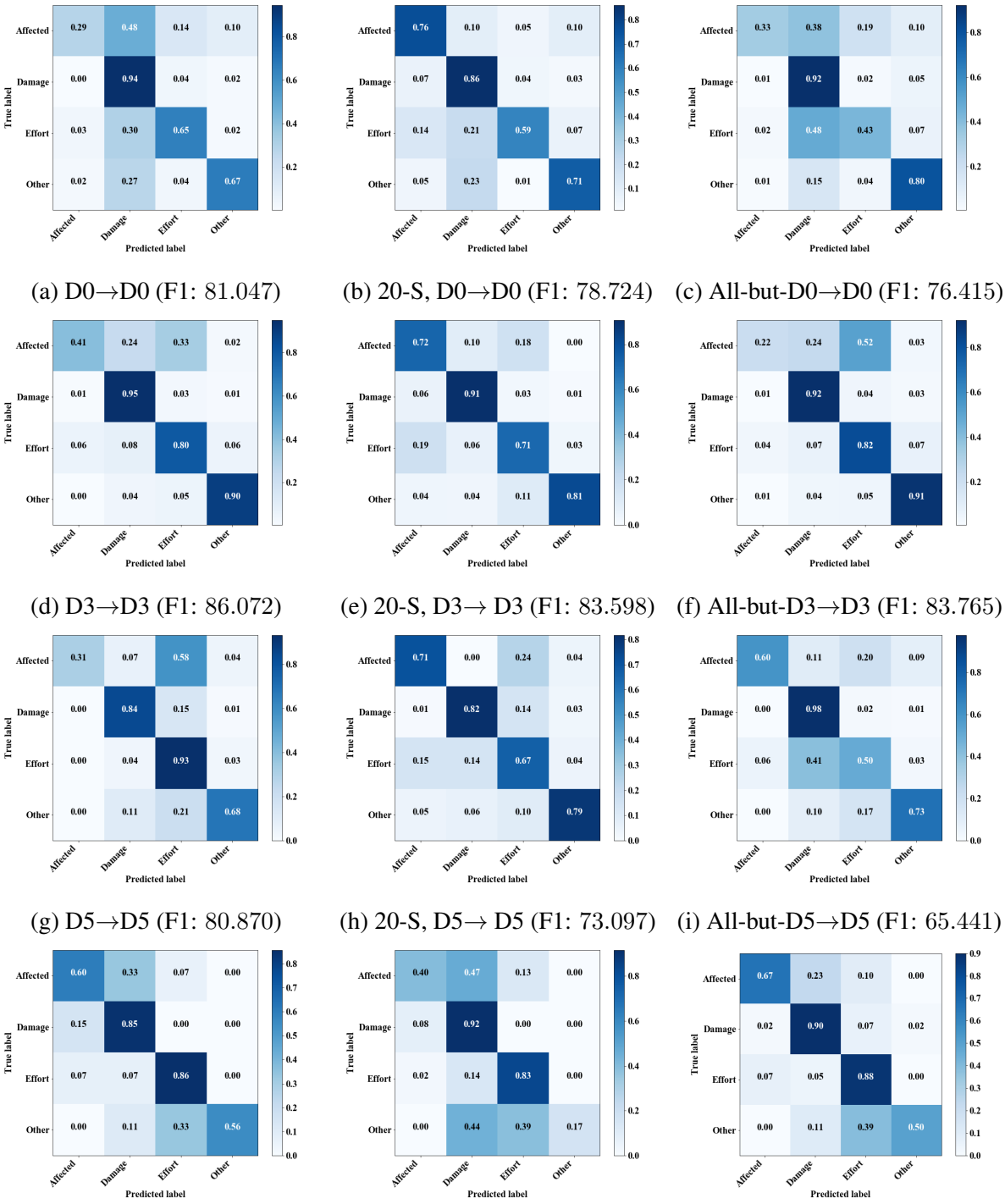
- While the *Damage* category has overall the largest number of false positives, it also has the largest number of true positives in all settings (according to the confusion matrices in Figure 3.3).

- Based on the confusion matrices, it is also interesting to see that in some cases (e.g., for the flood event), the patterns observed for supervised and all-but-one models are very similar , which shows that the diverse images in the all-but-one training subsets are sufficient to train effective models for a target event for which no labeled data is yet available.

3.7 Conclusion

We studied the use of advanced deep learning models, specifically transformer models, ViT/CSWin, and contrastive learning, CLIP, by comparison with a CNN-based model of ConvNeXts, for disaster image classification. Experimental results showed that CLIP outperformed the transformer models and also ConvNeXts in all settings for binary-class Informativeness and multi-class Humanitarian tasks. This suggests that transformer/contrastive learning models, pre-trained on large

amounts of data can capture diverse and complex patterns and can be potentially effective in disaster response efforts. Accurate and timely identification of disaster images provide useful situational awareness information and can greatly assist in response and recovery efforts. Overall, our study provides valuable insights into the potential of vision transformers and contrastive learning models for disaster response and highlights the importance of their usage in this field, to better support sustainable cities and communities and improve resilience. Future work can explore the use of other types of data, such as text, in conjunction with images to create multi-modal models for disaster response efforts.



(a) D0→D0 (F1: 81.047) (b) 20-S, D0→D0 (F1: 78.724) (c) All-but-D0→D0 (F1: 76.415)

(d) D3→D3 (F1: 86.072) (e) 20-S, D3→ D3 (F1: 83.598) (f) All-but-D3→D3 (F1: 83.765)

(g) D5→D5 (F1: 80.870) (h) 20-S, D5→ D5 (F1: 73.097) (i) All-but-D5→D5 (F1: 65.441)

(j) D6→D6 (F1: 76.243) (k) 20-S, D6→ D6 (F1: 65.997) (l) All-but-D6→D6 (F1: 79.034)

Figure 3.3: The confusion matrices of humanitarian tasks with D0 (Fire), D3 (Hurricane), D5 (Earthquake), and D6 (Flood) as the Target Domain, across different settings: *Event-specific supervised*, *Event-specific 20-shot (20-S)*, and *Cross-domain transfer (all-but-one)*

(a) (A,(A,A,A)) (b) (A,(E,A,A)) (c) (A,(D,A,D)) (d) (D,(D,E,D)) (e) (D,(E,E,E))

(f) (O,(O,O,O)) (g) (O,(O,A,O)) (h) (O,(E,E,E)) (i) (A,(A,A,A)) (j) (A,(E,E,E))

(k) (E,(E,E,E)) (l) (E,(E,A,E)) (m) (E,(E,E,D)) (n) (E,(D,D,D)) (o) (D,(D,D,D))

(p) (A,(A,A,A)) (q) (A,(A,D,A)) (r) (A,(D,D,A)) (s) (A,(D,D,D)) (t) (D,(D,D,D))

Figure 3.4: Error Analysis Images for Humanitarian Tasks with D0 (Fire), D3 (Hurricane), D5 (Earthquake), and D6 (Flood) as the target event: These figures illustrate classification instances correctly classified and misclassified in various settings, including *supervised*, *20-shot*, and *all-but-one*. In each row, the left side displays the name and type of the disaster. The captions for each image follow this format: (true label, (predicted by *supervised*, predicted by *20-shots*, predicted by *all-but-one*)). True labels are depicted in black, while predicted labels are shown in green for correct classifications and in red for misclassifications. The first letter of each class represents the labels: *Others (O)*, *Affected (A)*, *Effort (E)*, and *Damage (D)*.

Chapter 4

A comparison study for disaster tweet classification using deep learning models

4.1 Introduction

Social media plays a significant role in many people's lives and has changed the way people communicate virtually (Young et al., 2020). Being a faster way of news distribution and providing a two-way communication channel, social media has been increasingly gaining popularity and has grown into a main medium of human interaction and communication (Young et al., 2020). Therefore, it is not surprising that many people turn to social media during disaster events. (such as earthquakes, hurricanes, fires, etc.). They use social media platforms to connect with family and friends, to seek help and emotional support, to find information about food, shelter and transportation and to share and distribute information and news (Ahmed, 2011), especially as standard communication mediums (such as 911 lines) may be unavailable due to the large number of calls received during disasters (Villegas et al., 2018). For example, during the floods caused by Hurricane Harvey in Texas in August 2017, not being able to get through emergency lines, a lot of trapped people started posting messages on Twitter pleading for help. A baby was saved after the following message, along with a picture of a sleeping baby, was posted "Please help us she

is a newborn" (Koren, 2017). Such posts are of great importance for response organizations as they contain real-time information shared by eyewitnesses during ongoing disasters and can save lives (Roy et al., 2021).

The usefulness of information obtained from social media platforms for disaster response and management has been widely described in numerous works (Imran et al., 2020a). However, the data obtained from social media are noisy and contain irrelevant information, especially due to the large volume of tweets posted during emerging disasters (Alam et al., 2018c) and thus, the extraction and processing of relevant information is too difficult, time-consuming and costly to be manually conducted (Roy et al., 2021). Therefore, computational techniques are required to efficiently filter relevant posts, so that situational and actionable information can be acquired and used by emergency responders for helping the affected population (Alam et al., 2018c).

Some existing studies have utilized traditional machine learning (ML) techniques to process crisis-related social media data and help emergency response crews with extracting relevant posts out and use them to improve relief operations by identifying crisis relevant posts (Imran et al., 2018; Mazloom et al., 2019). More recent studies have used deep learning (DL) methods for this purpose (Roy et al., 2021; Alam et al., 2023; Khajwal et al., 2023; Sathishkumar et al., 2023; Li and Caragea, 2020; Kabir and Madria, 2019; Li et al., 2018; Taghian and Caragea, 2023b). In particular, the effectiveness of transformer-based models (e.g., BERT) in terms of classifying disaster tweets has been shown. For example, in a recent study (Alam et al., 2020b), BERT-type models had better performance as compared to convolutional neural networks (CNN) and FastText models (Joulin et al., 2016), and are currently considered to be state-of-the-art models for disaster tweet classification.

Another recent deep learning model which has also shown high potential in the text classification field is CapsNet model, first introduced for image classification (Sabour et al., 2017), which has been shown to outperform baseline models such as CNN, SVM-based models, and LSTM-based models in tasks such as sentiment classification (Wu et al., 2020; Zhang et al., 2018; Zhao et al., 2018), question categorization (Zhao et al., 2018), question answering (Zhao et al., 2019a),

news categorization (Zhao et al., 2018), and text classification (Chen et al., 2023). The superiority of CapsNet models has been demonstrated using various capsule architectures, including NLP-Capsule and CapsuleDAR (Zhang et al., 2018). In several studies, CapsNet models have exceeded baselines on multiple datasets and have proven their capability in transferring information from single-label to multi-label text classification (Zhao et al., 2018).

Given that CapsNet models have shown promising results for disaster image classification (Taghian and Caragea, 2021), due to their hierarchical structure preserving spatial relationships between features and given that they have also been effective for different text classification tasks, we aim to explore if using CapsNets can benefit crisis-related tweet classification tasks, as compared to BERT models and two other popular sequential models, LSTM and Bi-LSTM. Another motivation for this study was a lack of extensive comparison between CapsNets and state-of-the-art BERT models for disaster text classification, based on a careful literature review that we conducted. Therefore, we compared CapsNets with the state-of-the-art BERT model and also LSTM/Bi-LSTM models on the tasks of classifying crisis-related tweets in terms of both their informativeness with respect to disaster response (binary classification) and their humanitarian content (multi-class classification). For this purpose, we performed an extensive set of experiments using three datasets: CrisisLex, CrisisNLP, and CrisisBench, with the last one being a benchmark dataset consolidating eight prior datasets.

The rest of the chapter is structured as follows. Section 2 discusses the related work. Section 3 describes methods including LSTM/Bi-LSTM, CapsNets and their applications in NLP, and BERT. Sections 4 and 5 describe the implementation details (i.e., dataset, performance metrics, baselines, and experimental setup) and the result analysis, respectively. Finally, Section 6 concludes the chapter.

4.2 Related work

Text classification, one of the main tasks in natural language processing (NLP), has been widely explored in a variety of application domains (Kowsari et al., 2019). It has many time-critical applications, one of them being social media disaster message filtering and classification for disaster response.

Numerous traditional ML and NLP methods have been used to filter disaster-related tweets (Imran et al., 2018). Given that manually extracting the features required by ML methods is time-consuming and leads to loss of important textual elements, in recent years, DL approaches have been used to filter useful tweets for disaster response (Roy et al., 2021; Li and Caragea, 2020; Kabir and Madria, 2019). For example, Caragea et al. (2016) used CNNs to identify informative tweets posted during disaster events. Neppalli et al. (2018) performed a comparative study between deep learning models, such as CNNs and recurrent neural networks (RNNs). The authors showed that the DL models outperformed the standard ML classifiers on disaster-tweet classification even though only a small amount of labeled data was used for training the models. Furthermore, they showed that the CNN model outperformed the RNN model. Roy et al. (2021) proposed a hybrid CNN model that combines word and character-level embeddings to classify tweets into several disaster-related categories. Kabir and Madria (2019) designed a DL approach by combining attention-based Bi-directional Long Short-Term Memory (Bi-LSTM) networks with CNNs, and creating an auxiliary feature map to categorize the tweets.

Most recently, several studies have explored the use of transformer-based models for disaster tweets classification (Chanda, 2021; Ningsih and Hadiana, 2021; Zahera et al., 2019; Maharani, 2020; Naaz et al., 2021; Ray Chowdhury et al., 2020). For example, Ma (2019) showed that the BERT model outperformed the Bi-LSTM model but had the similar performance to customised BERT-based models that combined BERT with LSTM or CNN networks (i.e., BERT+LSTM and BERT+CNN). Alam et al. (2020b) compared three transformer-based models (BERT, DistilBERT, RoBERTa) with CNN and FastText models for both informativeness and humanitarian tasks on three datasets (one of them being the benchmark dataset, named CrisisBench). The au-

thors showed that the transformer-based models outperformed the CNN and FastText models, while the performance of the transformer-based models was similar. Given these results of prior studies, we consider BERT to be a state-of-the-art model for disaster tweet classification, and include it as a strong model in our study.

In the last few years, the CapsNet model proposed by Sabour et al. (2017) has gained a lot of attention in the area of image classification. Thanks to the dynamic routing procedure, CapsNets are capable of preserving spatial relationships between features, and can thus overcome one of the fundamental limitations of CNNs - the information loss caused by the use of pooling to achieve translation invariance (Patrick et al., 2022). While used in many application domains, CapsNet models have been explored in the context of disaster image classification with promising results (Taghian and Caragea, 2021). Given the results produced by CapsNets in image classification, Zhao et al. (2018) explored CapsNets with dynamic routing for text classifications and showed that they can achieve competitive results.

Following Zhao et al. (2018), many other authors have used the CapsNet model or its variants for the purpose of text classification in a variety of application domains (Zhao et al., 2019a; Wu et al., 2020; Zhang et al., 2018; Zhao et al., 2018; Demotte et al., 2021; Khikmatullaev, 2019; Yang et al., 2019; Kim et al., 2020). As some examples, Yang et al. (2019) proposed a cross-domain capsule network (TL-Capsule) and demonstrated its knowledge transfer capability for text classification, while Kim et al. (2020) incorporated an ELU-gate in the architecture of CapsNets and introduced a static routing procedure. A comparison between the static and dynamic routing on seven benchmark datasets showed the superiority of the static routing variant. Demotte et al. (2021) used shallow and deep CapsNets with both dynamic and static routing for social media content analysis and showed that shallow CapsNets with static routing outperformed deep CapsNets with either dynamic or static routing.

Despite the use of CapsNets for text classification and its proven superiority over CNN and RNN models, to the best of our knowledge, there is no prior study that extensively compares CapsNets with BERT-type models on text classification. Furthermore, CapsNets have not been

explored for disaster tweet classification, although they have shown promising results on disaster image classification (Taghian and Caragea, 2021). Therefore, the objective of our work was to perform a comprehensive comparison between CapsNet and BERT models on disaster tweet classification. We focused on CapsNets with both static and dynamic routings (Kim et al., 2020), among which static routing has given better results in prior works (Demotte et al., 2021; Kim et al., 2020), and addressed two specific tasks useful for disaster response: filtering tweets based on informativeness (a binary classification task) and identifying a tweet's humanitarian category (a multi-class classification task). We used three existing datasets in our experiments: CrisisLex, CrisisNLP, and CrisisBench, a benchmark dataset consolidating eight existing datasets (Alam et al., 2020b). In addition to the CapsNet and BERT models, we include LSTM/Bi-LSTM models in our study given that these models have been used as strong baselines for CapsNet and BERT models in prior works (Zhao et al., 2019a; Wu et al., 2020; Zhang et al., 2018; Zhao et al., 2018).

4.3 Methods

In this section, we provide background for the methods used in this study, including Long Short-Term Memory (LSTM) networks, Bidirectional LSTM (Bi-LSTM) networks, Bidirectional Encoder Representations from Transformers (BERT) networks, and Capsule Networks (CapsNets).

4.3.1 LSTM networks

Being able to process sequential data of arbitrary length, vanilla Recurrent Neural Networks (RNNs) have been extensively utilized in the filed of text classification (Zhao et al., 2019b). An RNN makes use of information from both previous and current words in order to learn sequential patterns (Minaee et al., 2021). LSTMs (Hochreiter and Schmidhuber, 1997) are RNN variants, which have the ability to learn long-term dependencies and have led to improved results in text classification (Liu and Guo, 2019). LSTMs use memory blocks to overcome the vanishing or exploding gradient problems faced by vanilla RNNs. More specifically, an LSTM unit includes a

memory cell and a forget gate, together with input and output gates. The memory cell collects and remembers information, while the forget gate decides when and how much of the old information to forget. The input and output gates are in charge of controlling the amount of information that goes into and out of the cell, respectively (Hochreiter and Schmidhuber, 1997; Liu and Guo, 2019).

4.3.2 Bidirectional LSTM (Bi-LSTM)

Processing data only in one direction (utilizing only the past information and not being able to access the future one), standard LSTM models might suffer from the drawback of misinterpreting the context or not fully understanding it (Liu and Guo, 2019). On the contrary, Bi-LSTM (Graves and Schmidhuber, 2005), an LSTM variant, can process information in a bidirectional way, accessing and analysing both past and future words simultaneously and combining them for current output prediction (Lu et al., 2020). In fact, a Bi-LSTM model uses two hidden LSTM layers in its architecture, one forward and one backward, then combines the outputs of those two layers to capture both the forward and backward data streams in a sequence (Liu and Guo, 2019).

4.3.3 Bidirectional encoder representations from transformers (BERT)

The Transformer model, which makes use of self-attention (Vaswani et al., 2017), has achieved outstanding performance in many NLP applications (Acheampong et al., 2021). Thanks to the attention mechanism, transformers can process the words in a sequence simultaneously, leading to a more computationally efficient implementation which makes it possible to train models on very large corpora (Vaswani et al., 2017). With the emergence of the transformer, different transformer-based models, including BERT, have been developed for text encoding and classification tasks. BERT (Devlin et al., 2018) is based on a multilayered bidirectional transformer encoder, as opposed to prior contextual word embedding models which are mainly unidirectional (e.g., Elmo) (Kaliyar, 2020). The bidirectional feature allows BERT models to achieve state-of-

the-art performances on different NLP-based tasks (Kaliyar, 2020).

Devlin et al. (2018) provides two BERT structures including BERT-Base and BERT-Large. The former has 12 layers (i.e., transformer blocks) with 12 self-attention heads, hidden layer size of 768 and total parameters of 110 millions, while the latter has 24 layers (i.e., transformer blocks) with 16 self-attention heads, hidden layer size of 1024 and total parameters of 340 millions.

There are different variants of BERT such as RoBERTa (Robustly Optimized BERT pre-training Approach) proposed by Liu et al. (2019), and DistilBERT proposed by Sanh et al. (2019), which have been previously used for disaster tweet classification, in addition to BERT. As mentioned before, Ma (2019) showed that these three models have similar performance in terms of informativeness and humanitarian category classification tasks. Therefore, we chose BERT as one of our baselines.

4.3.4 Capsule networks

CapsNets were first introduced in the field of image classification (Sabour et al., 2017) to address CNNs limitations, mainly their loss of information due to using subsampling/pooling layers, leading to translational invariance where neurons' activity is invariant to viewpoint translation. Translational invariance is the cause of the "Picasso problem" - CNNs trying to identify an image by only detecting its components without considering the spatial relationship between those components. On the other hand, CapsNets make use of the equivariant property (meaning that neurons' activity changes as viewpoint changes), and thus, preserve the spatial relationship between components (Patrick et al., 2022).

A CapsNet contains multiple layers of capsules, i.e., groups of neurons each of which can capture a different property in an object (such as its position, size, etc.). This enables a capsule to have inputs and outputs in the form of a vector (whose length determines the likelihood of an object) or a matrix (together with an activation probability). However, a single neuron in a neural network can only have scalar values as inputs and outputs (Sabour et al., 2017; Hinton et al., 2018).

A routing-by-agreement algorithm is used to determine the part-whole relationships between lower-layer capsules (parts) and higher-layer capsules (wholes). According to this algorithm, if several predictions made by active lower-layer capsules (through transformation matrices) agree, a higher-layer capsule becomes active (Sabour et al., 2017). For the first time, Zhao et al. (2018) demonstrated the effectiveness of CapsNets for text classification. For the purpose of stabilizing the dynamic routing process and mitigating the noise caused by capsules not being effectively trained or by capsule with "background" information of the input sequence (e.g., stop words and unrelated words with respect to specific categories), they proposed three strategies to enhance the performance of dynamic routing algorithm. The first strategy was to include an extra "Orphan" category to CapsNet with the purpose of capturing the "background" knowledge. This extra category causes CapsNet to more efficiently learn the part-whole relationship. The second strategy was to use Leaky-softmax instead of standard softmax for calculating connection strength between the part-whole capsules. The third strategy was to use lower-layer capsule's probability of existence for iteratively updating the connection strength.

However, more recently, Kim et al. (2020) proposed a static routing for text classification tasks and showed its superiority over dynamic routing. Therefore, in this study, we also use the CapsNet architecture with static routing as proposed by Kim et al. (2020) and compare it with the CapsNet model which uses dynamic routing. Moreover, the architecture in (Kim et al., 2020) has an added ELU-gate. Compared to the original CapsNet structure, the benefit of which is the distribution of relevant information through the words and tokens in a given input sequence, this gate decides whether to activate a feature or not. The advantage of the ELU-gate unit over pooling is that it would preserve spatial information.

4.4 Implementation details

In this section, the datasets used in the experiments, along with the experimental setup, and evaluation metrics are presented. The purpose of the conducted experiments is to answer the following

research questions which motivated this study:

- How do the CapsNet models compare to state-of-the-art BERT models and to the popular LSTM/Bi-LSTM models when used for classifying crisis-related tweets in terms of informativeness (binary classification) and humanitarian category (multi-class classification)?
- How do the two routing algorithms used in CapsNets compare to each other when classifying crisis-related tweets in terms of informativeness and humanitarian category?

4.4.1 Datasets

The datasets used in this study include CrisisLex, CrisisNLP, and CrisisBench, the statistics of which can be found in Tables 4.1, 4.2, and 4.3, respectively. All three datasets have annotations for both informativeness and humanitarian category tasks. For all three datasets, the informativeness task includes two classes of Informative and Not-informative; however, for the humanitarian category task, there are 6, 10 and 11 classes in CrisisLex, CrisisNLP and CrisisBench datasets, respectively, as shown in the corresponding tables. A brief description of each dataset is given below.

- **CrisisLex** is a combination of two datasets, specifically, CrisisLexT26 and CrisisLexT6 (Olteanu et al., 2014). The CrisisLexT26 dataset consists of tweets from 26 disasters which happened during 2012 and 2013, while the CrisisLexT6 dataset consists of tweets from 6 disasters which happened between October 2012 and July 2013 (Alam et al., 2020b).
- The **CrisisNLP** dataset includes tweets from 19 disasters that occurred between years 2013 and 2015 (Imran et al., 2016a).
- **CrisisBench** is a benchmark dataset constructed by Alam et al. (2020b). This dataset is a consolidation of eight prior datasets, specifically, CrisisLex, CrisisNLP, SWDM2013, ISCRAM2013, Disaster Response Data (DRD), Disasters on Social Media (DSM), CrisisMMD, and AIDR.

Informativeness	**Train**	**Dev**	**Test**	**Total**
Informative	25955	3727	7447	37129
Not informative	18862	2769	5384	27015
Total	44817	6496	12831	64144
Humanitarian	**Train**	**Dev**	**Test**	**Total**
Affected individual	2125	317	601	3043
Caution and advice	912	143	247	1302
Donation and volunteering	1031	158	293	1482
Infrastructure and utilities damage	752	79	216	1047
Not humanitarian	18844	2748	5423	27015
Sympathy and support	1800	283	532	2615
Total	25464	3728	7312	36504

Table 4.1: Statistics for CrisisLex dataset

Informativeness	**Train**	**Dev**	**Test**	**Total**
Informative	14865	2182	4087	21134
Not informative	11298	1645	3119	16062
Total	26163	3827	7206	37196
Humanitarian	**Train**	**Dev**	**Test**	**Total**
Caution and advice	706	99	198	1003
Displaced and evacuations	322	51	88	461
Donation and volunteering	1647	265	470	2382
Infrastructure and utilities damage	1128	163	300	1591
Injured or dead people	1235	165	362	1762
Missing and found people	278	44	80	402
Not humanitarian	11227	1686	3150	16063
Requests or needs	103	19	29	151
Response efforts	780	113	221	1114
Sympathy and support	1624	237	455	2316
Total	19050	2842	5353	27245

Table 4.2: Statistics for CrisisNLP dataset

4.4.2 Performance metrics

In the experiments, we compare the CapsNet models (both with static and dynamic routing algorithms, shown by CapsNet-s and CapsNet-d, respectively) with BERT models and also with LSTM

Informativeness	**Train**	**Dev**	**Test**	**Total**
Informative	65826	9594	18626	94046
Not informative	43970	6414	12469	62853
Total	109796	16008	31095	156899
Humanitarian	**Train**	**Dev**	**Test**	**Total**
Affected individual	2454	367	693	3514
Caution and advice	2101	309	583	2993
Displaced and evacuations	359	53	99	511
Donation and volunteering	5184	763	1453	7400
Infrastructure and utilities damage	3541	511	1004	5056
Injured or dead people	1945	271	561	2777
Missing and found people	373	55	103	531
Not humanitarian	36109	5270	10256	51635
Requests or needs	4840	705	1372	6917
Response efforts	780	113	221	1114
Sympathy and support	3549	540	1020	5109
Total	61235	8957	17365	87557

Table 4.3: Statistics for CrisisBench dataset

and Bi-LSTM models on both binary (tweet informativeness) and multi-class (tweet humanitarian category) classification tasks. For both types of tasks, the comparison is performed using several standard evaluation metrics, specifically, Precision, Recall and F1 scores for each class and Weighted Precision, Weighted Recall and Weighted F1 scores for the overall performance of each model.

4.4.3 Experimental setup

Since train/test/dev subsets are provided for the three datasets used in the study, the experiments were conducted three times using three different seeds and the average of the three runs was reported in the tables.

The hyperparameters employed in the experiments were obtained through fine-tuning on the respective development (dev) datasets. Specifically, for the LSTM and Bi-LSTM experiments, the following hyperparameters were utilized: hidden size of 256, maximum sequence length of 60,

learning rate of 5e-4, drop ratio of 0.1, weight decay of 0 (i.e., no L2 regularization), a batch size of 40, maximum number of epochs of 10.

For the BERT experiments, BERT base uncased model, which has 12 layers (i.e., Transformer blocks), 12 self-attention heads, and with a hidden size of 768, was used. The hyperparameters used in these experiments are: maximum sequence length of 60, learning rate of 2e-5, drop ratio of 0, weight decay of 0, a batch size of 40, maximum number of epochs of 10.

The hyperparameters used in the CapsNet with static routing experiments are: maximum sequence length of 60, learning rate of 5e-5, drop ratio of 0.6, weight decay of 0, a batch size of 40, maximum number of epochs of 50. The hyperparameters used in the CapsNet with dynamic routing experiments are: Maximum sequence length of 60, learning rate of 4e-5, drop ratio of 0.6, weight decay of 1e-5, a batch size of 40, maximum number of epochs of 50.

Adam optimizer was used for all the experiments. Furthermore, for all experiments, the number of epochs resulting in the best accuracy on the development set was used for evaluating performance on the test set.

4.5 Experimental results and discussions

The experimental results of crisis tweet classification for informativeness category task of all three datasets are presented in Table 4.4 and for humanitarian category task of CrisisLex, CrisisNLP, and CrisisBench are presented in Tables 4.5, 4.6, and 4.7, respectively. The tables depict the performance of the models for each class (namely, Precision, Recall, and F1-score) in the datasets, along with the overall performance of the models in the "Overall" row, which shows the Weighted Precision, Weighted Recall, and Weighted F1-score across all classes of each dataset. We discuss the answers to our questions using these results in the following paragraphs.

Our first research question was to compare CapsNet models to state-of-the-art BERT models and also to LSTM/Bi-LSTM models trained to classify crisis-related tweets in terms of informativeness (binary classification) and humanitarian category (multi-class classification) using three

Dataset	Classes	Metrics	Models				
			LSTM	**Bi-LSTM**	**Bert**	**CapsNet_s**	**CapsNet_d**
CrisisLex	**Informative**	Precision	94.308	**94.746**	94.255	94.469	94.334
		Recall	96.258	95.815	**96.567**	95.108	95.609
		F1-score	95.273	95.275	**95.393**	94.785	94.967
	Not-informative	Precision	94.671	94.127	**95.101**	93.175	93.812
		Recall	91.961	**92.642**	91.846	92.292	92.057
		F1-score	93.296	93.374	**93.436**	92.728	92.926
	Overall	Precision	94.460	94.486	**94.610**	93.926	94.115
		Recall	94.455	94.484	**94.586**	93.926	94.119
		F1-score	94.443	94.477	**94.571**	93.922	94.110
CrisisNLP	**Informative**	Precision	84.249	85.099	**88.877**	84.075	84.270
		Recall	**87.342**	86.747	85.825	85.735	86.184
		F1-score	85.764	85.879	**87.297**	84.888	85.202
	Not-informative	Precision	**82.586**	82.256	82.279	80.823	81.372
		Recall	78.593	79.994	**85.861**	78.690	78.882
		F1-score	80.533	81.042	**83.993**	79.725	80.083
	Overall	Precision	83.530	83.868	**86.022**	82.667	83.015
		Recall	83.556	83.824	**85.840**	82.686	83.023
		F1-score	83.500	83.786	**85.867**	82.654	82.986
CrisisBench	**Informative**	Precision	88.919	88.036	**89.452**	87.916	86.830
		Recall	89.777	90.543	**90.708**	89.622	91.168
		F1-score	89.336	89.270	**90.057**	88.755	88.944
	Not-informative	Precision	84.526	85.244	85.872	84.041	**85.740**
		Recall	83.243	81.601	**83.956**	81.572	79.321
		F1-score	83.857	83.377	**84.862**	82.776	82.399
	Overall	Precision	87.158	86.917	**88.016**	86.362	86.393
		Recall	87.158	86.959	**88.002**	86.395	86.419
		F1-score	87.140	86.908	**87.975**	86.358	86.320

Table 4.4: Experiment results for tweet informativeness task for all three datasets, for each class and the overall weighted metrics, with the best value of each metric displayed in bold within each row

datasets, specifically CrisisLex, CrisisNLP, and CrisisBench. For each task, we initially analyze the overall performance of the models and then examine the results for each class in the dataset. As can be seen in Tables 4.4, 4.5, 4.6, and 4.7 for both tasks, BERT models are the best models in terms of all the evaluation metrics considered, when compared to LSTM, Bi-LSTM, CapsNet-s and Capsnet-d models.

For the tweet informativeness task, the F1-scores of BERT models are 94.571, 85.867, and 87.975 for CrisisLex, CrisisNLP, and CrisisBench datasets, respectively. These scores are higher than those of the other four models for the corresponding datasets. For this classification task and for the two datasets of CrisisLex, CrisisNLP, the order of other models, from best to worst, is

Classes	Metrics	Models				
		LSTM	Bi-LSTM	Bert	CapsNet_s	CapsNet_d
Affected individual	Precision	**85.366**	83.053	**85.221**	82.065	83.216
	Recall	78.591	81.697	**86.023**	83.250	82.917
	F1-score	81.727	82.358	**85.603**	82.632	83.041
Caution and advice	Precision	64.732	66.500	**72.484**	67.972	66.109
	Recall	64.777	59.784	**71.660**	64.642	66.801
	F1-score	64.639	62.956	**72.042**	66.257	66.388
Donation and volunteering	Precision	76.334	76.660	**81.233**	78.492	77.573
	Recall	**79.181**	76.678	77.702	75.313	76.451
	F1-score	77.650	76.624	**79.408**	76.832	76.987
Infrastructure and utilities damage	Precision	55.944	56.238	**66.348**	61.407	65.504
	Recall	**67.747**	62.037	58.642	46.450	43.364
	F1-score	61.112	58.378	**62.140**	52.864	52.158
Not humanitarian	Precision	**97.910**	97.308	97.543	96.555	96.239
	Recall	97.554	97.855	97.861	97.984	**98.359**
	F1-score	**97.731**	97.577	97.702	97.264	97.286
Sympathy and support	Precision	79.863	81.578	**82.079**	79.516	81.764
	Recall	80.827	77.882	**84.586**	77.506	75.000
	F1-score	80.330	79.581	**83.276**	78.464	78.232
Overall	Precision	92.341	91.911	**92.984**	91.500	91.442
	Recall	92.054	91.881	**93.071**	91.808	91.822
	F1-score	92.145	91.850	**93.008**	91.609	91.539

Table 4.5: Experiment results for tweet humanitarian task for the CrisisLex datasets, for each class and the overall weighted metrics, with the best value of each metric displayed in bold within each row

Bi-LSTM, LSTM, CapsNet-d and CapsNet-s (with F1-scores of 94.477, 94.443, 94.110, 93.922, respectively, for CrisisLex dataset and 83.786, 83.500, 82.986, 82.654, respectively, for CrisisNLP dataset). For the CrisisBench dataset, the order is LSTM, Bi-LSTM, CapsNet-s and CapsNet-d (with F1-scores of 87.140, 86.908, 86.358, and 86.320, respectively). As for the performance on the classes of this task, we observe a similar behavior, with BERT model still having the highest F1-score for both classes and for all three datasets (with F1-score of 95.393, 87.297, and 90.057 for Informative class and 93.436, 83.993, 84.862 for Not-informative class for CrisisLex, CrisisNLP, and CrisisBench datasets, respectively). LSTM and BiLSTM models typically rank in the second or third positions (with comparable F1-scores), while CapsNet-s and CapsNet-d usually occupy the third or fourth positions (with similarly close F1-scores).

For the crisis tweet humanitarian category, the F1-scores of the BERT models are 93.008, 81.300, and 86.708 for CrisisLex, CrisisNLP, and CrisisBench datasets, respectively. As for the

Classes	Metrics	Models				
		LSTM	**Bi-LSTM**	**Bert**	**CapsNet_s**	**CapsNet_d**
Caution and advice	Precision	74.861	69.971	**77.631**	66.060	70.450
	Recall	52.188	**62.626**	57.407	56.061	50.673
	F1-score	61.410	65.862	**65.869**	60.364	58.713
Displaced and evacuations	Precision	52.648	55.889	**64.369**	61.226	33.492
	Recall	40.530	41.288	**59.470**	29.925	18.939
	F1-score	45.375	46.983	**61.188**	39.650	24.019
Donation and volunteering	Precision	73.526	61.887	**78.905**	64.779	62.500
	Recall	65.816	**77.873**	73.688	70.142	71.986
	F1-score	69.205	68.824	**76.007**	67.285	66.708
Infrastructure and utilities damage	Precision	69.048	75.072	**76.910**	76.248	63.742
	Recall	65.222	62.667	**67.000**	61.667	68.222
	F1-score	67.029	67.970	**71.473**	67.924	65.743
Injured or dead people	Precision	82.086	83.200	**86.100**	82.294	82.242
	Recall	88.030	86.464	**88.674**	80.019	79.374
	F1-score	84.944	84.785	**87.328**	81.093	80.662
Missing and found people	Precision	44.868	44.762	**55.407**	54.939	54.550
	Recall	39.583	39.583	**42.083**	33.750	21.250
	F1-score	41.956	41.949	**47.798**	41.606	28.776
Not humanitarian	Precision	83.013	84.561	**85.820**	81.686	81.529
	Recall	92.688	91.566	**94.656**	92.857	92.222
	F1-score	87.575	87.923	**90.005**	86.912	86.526
Requests or needs	Precision	13.333	2.564	**45.707**	0.000	0.000
	Recall	2.299	1.149	**13.793**	0.000	0.000
	F1-score	3.922	1.587	**20.880**	0.000	0.000
Response efforts	Precision	38.067	40.133	**49.081**	42.209	32.486
	Recall	17.949	17.496	**37.255**	8.145	6.033
	F1-score	23.549	23.719	**41.376**	12.472	9.932
Sympathy and support	Precision	68.210	74.843	**81.522**	63.531	65.120
	Recall	52.349	48.164	**58.462**	49.084	47.692
	F1-score	59.196	58.085	**67.925**	55.273	54.961
Overall	Precision	76.474	77.237	**81.523**	75.272	73.446
	Recall	78.326	78.494	**82.234**	77.544	76.586
	F1-score	76.784	77.089	**81.300**	75.302	73.991

Table 4.6: Experiment results for tweet humanitarian task for the CrisisNLP datasets, for each class and the overall weighted metrics, with the best value of each metric displayed in bold within each row

crisis tweet informativeness tasks, these scores are higher than those of the other four models for the corresponding datasets. For this classification task and for the dataset of CrisisLex, the order of other models, is LSTM, Bi-LSTM, CapsNet-s and CapsNet-d (with F1-scores of 92.145, 91.850, 91.609, and 91.539, respectively), while for the other two datasets, the order is Bi-LSTM, LSTM, CapsNet-s and CapsNet-d (with F1-scores of 77.089, 76.784, 75.302, 73.991, respectively,

Classes	Metrics	Models				
		LSTM	Bi-LSTM	Bert	CapsNet_s	CapsNet_d
Affected individual	Precision	78.190	75.118	80.910	**81.212**	77.536
	Recall	74.422	76.975	**78.372**	66.281	68.160
	F1-score	76.105	75.838	**79.587**	72.931	72.509
Caution and advice	Precision	67.865	62.683	**70.034**	66.864	64.088
	Recall	59.023	66.724	**70.783**	57.126	59.713
	F1-score	63.041	64.445	**70.349**	61.605	61.702
Displaced and evacuations	Precision	52.226	50.211	**64.992**	29.580	41.667
	Recall	35.690	38.384	**55.892**	7.744	2.020
	F1-score	42.206	42.958	**59.485**	12.272	3.639
Donation and volunteering	Precision	73.573	76.623	**76.708**	70.138	68.442
	Recall	77.135	74.885	**82.048**	78.007	79.270
	F1-score	75.308	75.637	**79.254**	73.860	73.456
Infrastructure and utilities damage	Precision	73.320	66.841	**73.555**	65.263	68.253
	Recall	58.634	65.930	**67.864**	63.554	60.576
	F1-score	64.944	66.240	**70.518**	64.367	64.125
Injured or dead people	Precision	82.823	82.616	**82.860**	78.888	76.657
	Recall	78.295	80.083	**84.551**	74.120	76.923
	F1-score	80.450	81.323	**83.572**	76.420	76.731
Missing and found people	Precision	54.618	47.643	**63.193**	39.958	70.000
	Recall	29.126	38.511	**39.482**	9.385	1.618
	F1-score	36.850	42.509	**48.595**	14.259	3.128
Not humanitarian	Precision	88.379	90.170	**91.615**	87.765	88.190
	Recall	**94.804**	93.639	94.535	94.258	94.081
	F1-score	91.476	91.865	**93.051**	90.892	91.029
Requests or needs	Precision	89.681	90.957	**94.112**	87.298	89.368
	Recall	85.359	85.017	**90.414**	82.748	81.552
	F1-score	87.269	87.833	**92.220**	84.868	85.274
Response efforts	Precision	35.116	36.993	**48.832**	13.889	0.000
	Recall	18.703	16.139	**30.317**	0.754	0.000
	F1-score	23.226	21.325	**37.374**	1.431	0.000
Sympathy and support	Precision	80.494	75.628	**82.361**	75.047	73.356
	Recall	60.713	64.475	**69.085**	59.739	62.091
	F1-score	69.171	69.268	**75.140**	66.307	67.199
Overall	Precision	83.536	83.956	**86.679**	81.381	81.483
	Recall	84.217	84.413	**87.000**	82.953	82.992
	F1-score	83.528	83.988	**86.708**	81.778	81.757

Table 4.7: Experiment results for tweet humanitarian task for the CrisisNLP datasets, for each class and the overall weighted metrics, with the best value of each metric displayed in bold within each row

for CrisisNLP dataset and 83.988, 83.528, 81.778, 81.757, respectively, for CrisisBench dataset). BERT outperforms other models in terms of F1-scores on all classes in CrisisNLP and CrisisBench datasets. In CrisisLex, LSTM leads only in the "Not humanitarian" class with an F1-score of 97.731, while BERT is close behind with a score of 97.702. For the rest of the classes in

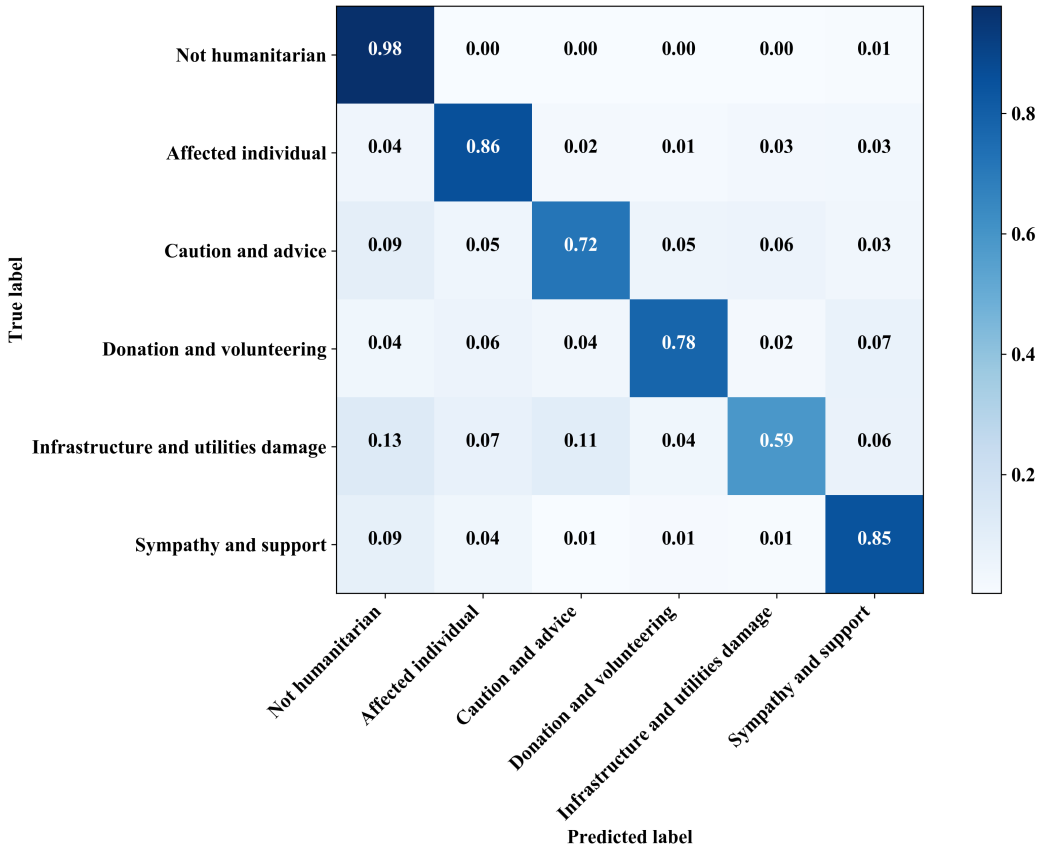


Figure 4.1: BERT's CM for humanitarian task on CrisisLex

CrisisLex, BERT still outperforms other models. BERT's classifications of instances in different classes can be better understood through the confusion matrices (CM) in Figures 4.1, 4.2, and 4.3 for the three datasets. As can be seen in these Figures, for the CrisisLex dataset, the model seems to have difficulties in correctly identifying instances that relate to "Infrastructure and utilities damage", with the accuracy of 0.59. For CrisisNLP, most of the instances of "Requests or needs" and "Response efforts" classes (with the lowest accuracy of 0.14 and 0.37, respectively) have been mis-classified as "Not humanitarian". Similarly, in CrisisBench, most of the instances of "Response efforts" class (with the lowest accuracy of 0.30) have been mis-classified as "Not humanitarian". Based on these results, we can conclude that BERT is the best model for classifying crisis-related tweet datasets, while CapsNet models (with both static and dynamic routing)

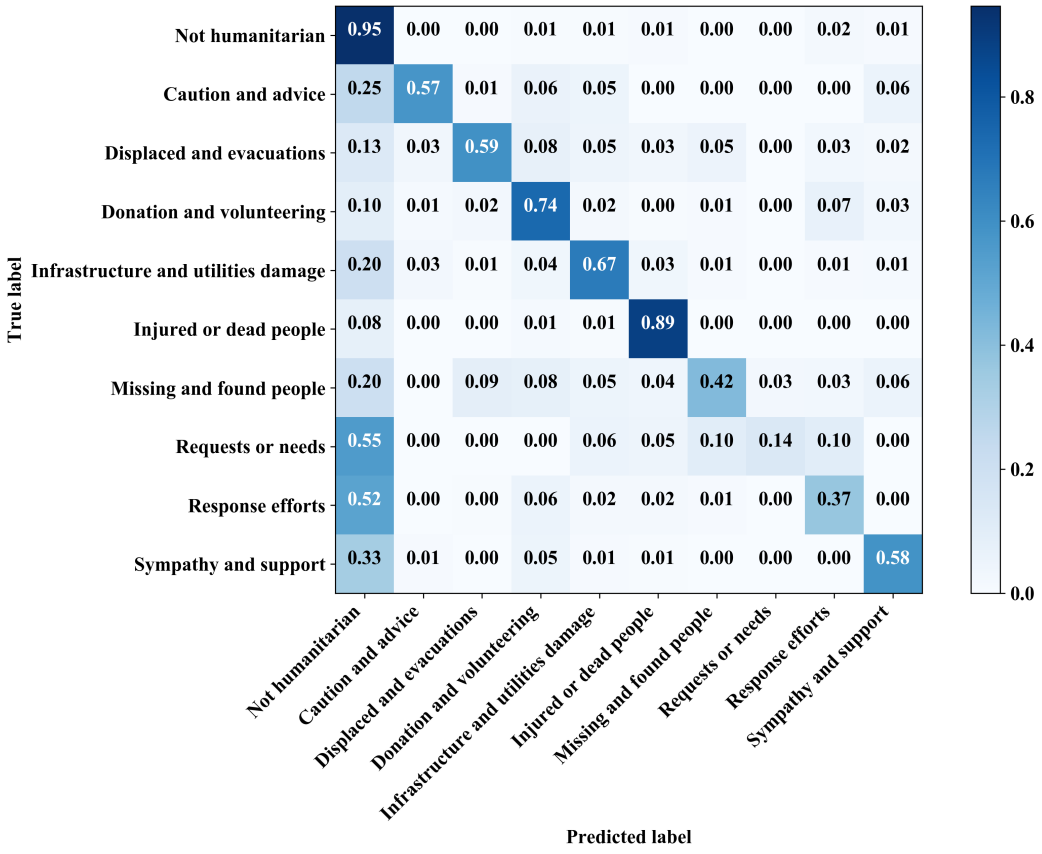


Figure 4.2: BERT's CM for humanitarian task on CrisisNLP

have the worst performance. Therefore, the properties of CapsNet do not seem to benefit much the task of classifying crisis-related tweet datasets.

Our second research question was to compare two routing algorithms used in CapsNet models, mainly static and dynamic routings, for classifying crisis-related tweets in terms of informativeness and humanitarian category using the three datasets considered in our study. For the crisis tweet informativeness task and for two datasets (CrisisLex, CrisisNLP), the performance of CapsNet-d is better than that of CapsNet-s (with F1-scores of 94.110, 93.922, respectively, for the CrisisLex dataset, and 82.986, and 82.654, respectively, for the CrisisNLP dataset). For the CrisisBench dataset, the F1-score of CapsNet-s (86.358) is higher than that of CapsNet-d (86.320). However, for the tweet humanitarian category task, for all three datasets, CapsNet-s outperforms

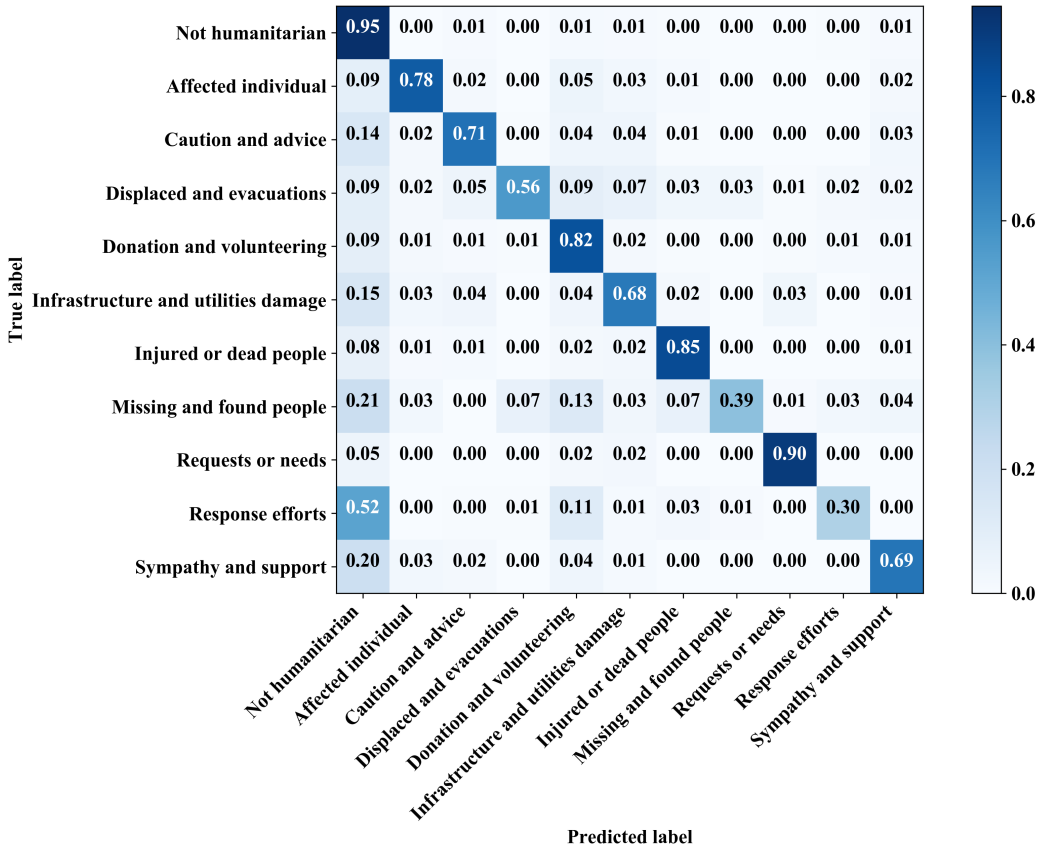


Figure 4.3: BERT's CM for humanitarian task on CrisisBench

CapsNet-d, with F1-scores of 91.609 versus 91.539, respectively, for CrisisLex, 75.302 versus 73.991, respectively, for CrisisNLP dataset, and 81.778 versus 81.757, respectively, for the CrisisBench dataset. However, in all cases, the F1-scores of the two routing algorithms are pretty close to each other. Based on the results, we can conclude that when classifying crisis-related tweets both in terms of informativeness (binary classification) and humanitarian category (multi-class classification), both static routing and dynamic routing algorithms have similar performance

Overall, based on the obtained results, BERT emerges as a stronger candidate compared to CapsNet for disaster tweet classifications. The superior performance of BERT can be attributed to several key factors. Firstly, BERT's pre-training on extensive data enables it to acquire a deep understanding of complex patterns and contextual information. Additionally, its utilization of

the attention mechanisms and transformer architecture empowers BERT to capture long-range dependencies in language, resulting in a better comprehension and generation of natural language. While CapsNet shows promise in specific image-related tasks, thanks to its equivariant property that preserves spatial relationships between components, it has not attained the same level of success as BERT in the field of natural language processing.

4.6 Conclusions and future work

In this study, we compared CapsNets (with static and dynamic routing algorithms) with BERT models, as well as LSTM/Bi-LSTM on crisis-related tweet classification tasks, specifically tweet informativeness (binary classification), and tweet humanitarian category (multi-class classification), using three datasets, specifically, CrisisLex, CrisisNLP, and CrisisBench. The results show that the BERT models have the best performance for classifying crisis-related data, while CapsNet models (with both static and dynamic routing) have the worst performance. Also, for both Informativeness task and Humanitarian task, both static routing and dynamic routing would result in a similar performance.

For future work, based on the superior performance of the BERT model over the CapsNet model, our intention is to explore the potential of more recent transformer-based models for disaster tweet classification. Additionally, we aim to investigate the effectiveness of transformer-based architectures in disaster image classification. Furthermore, we plan to apply these models to a dataset that contains both textual and visual modalities, with the objective of developing a multimodal model based on transformer-based architectures for classifying disaster-related posts, since sometimes the information conveyed in a text and its corresponding image can be complementary, and leveraging both modalities simultaneously may lead to enhanced performance.

Chapter 5

Exploring the potential of large language models for disaster tweet classification in zero-shot and few-shot settings

5.1 Introduction

The aftermath of a disaster is a critical time phase where obtaining timely and accurate information is paramount for effective response and recovery efforts. Rapid assessment of the situation, including identifying informative tweets and their categories, plays a vital role in directing resources and prioritizing rescue operations (Alam et al., 2020a). This is particularly important because Twitter frequently serves as a central platform for real-time updates, eyewitness accounts, and requests for assistance during such disaster events (King, 2018). However, traditional tweet classification techniques often struggle in these scenarios due to the limited availability of labeled training data specific to disaster situations right after the disaster. Gathering large, labeled datasets for disaster image classification presents a significant challenge, as obtaining high-quality labeled data often requires manual annotation by experts, a time-consuming and resource-intensive process. This holds especially true in the immediate aftermath of a disaster, when rapid response is

essential. The limited availability of labeled training data significantly hinders the performance of traditional machine learning models, which rely heavily on large datasets for effective training.

One approach to address this challenge involves utilizing domain adaptation techniques (Li et al., 2018, 2019, 2015; Imran et al., 2016b). These methods aim to improve model performance on a new disaster data (target domain) that differs from the training data (source domain) in terms of disaster type, location, and time. However, with the rise of Large Language Models (LLMs), there is a new way of dealing with this situation. Large Language Models have emerged as powerful tools in various artificial intelligence applications, particularly those characterized by limited training data. These models are trained on massive amounts of text data (Touvron et al., 2023a), enabling them to learn complex relationships between words and concepts. This extensive pretraining allows LLMs to generalize to new tasks with minimal fine-tuning, a characteristic that makes them well-suited for disaster tweet classification.

In fact, the swift progress of Large Language Models (LLMs) in recent years, has revolutionized natural language processing. These models exhibit impressive capabilities across various tasks, mainly due to their exceptional zero-shot and few-shot generalization. For example, ChatGPT, based on GPT-3.5 and GPT-4, has demonstrated robust dialogue skills (Xie et al., 2023). Models like GPT-3, GPT-4, and Meta's Llama 2 (Touvron et al., 2023a) are particularly adept at handling tasks with little to no training data, marking a significant advancement over traditional machine learning models that require extensive datasets for effective training. This remarkable generalization ability is especially valuable in scenarios such as disaster response, where obtaining labeled training data quickly is often difficult. Thus, these models excel in situations with limited labeled data, as is frequently the case immediately after a disaster. By leveraging their extensive pretraining, LLMs can adapt to new tasks rapidly and effectively, making them an essential tool for quick response efforts.

In this chapter, we utilize Llama 2 and ChatGPT as two Large Language Models (LLMs) in zero-shot, one-shot, and (for Llama 2) five-shot learning settings. Our focus is on tasks related to informativeness and humanitarian content, using the same three datasets as in the previous chapter.

Specifically, we aim to compare the performance of these LLMs against fine-tuned models from the previous chapter. As fine-tuned models, we employ BERT (which showed the best overall results in the previous chapter), CapsNet with static routing (given that both static and dynamic routing yielded comparable results previously), and Bi-LSTM (which performed slightly better than the standard LSTM). We do not present the numeric results for the fine-tuned models in this chapter since they were detailed in the previous chapter. Instead, we present the numeric results of the LLMs in tables, along with graphs showing the F1-scores of both LLMs and the three selected fine-tuned models from the previous chapter for convenient comparison.

The contributions of this chapter are as follows:

- We utilized ChatGPT and Llama 2 in zero-shot and few-shot learning settings.
- We conducted experiments with Llama 2 in zero-shot, one-shot, and five-shot settings, and with ChatGPT in zero-shot and one-shot settings for both tasks.
- We compared the results of these models with three selected fine-tuned models from the previous chapter, namely BERT (the best performer), CapsNet with static routing, and Bi-LSTM, by including their F1 score comparisons.

These contributions highlight the potential of LLMs in scenarios with limited labeled data, which is often the case immediately following a disaster, providing a valuable tool for rapid response efforts. Our extended comparison reveals that the fine-tuned BERT model remains the best model for the tasks considered. However, ChatGPT and Llama 2 (esp. ChatGPT) exhibit very promising results in low-resource settings typical of practical scenarios in the crisis domain. ChatGPT, delivered good performance even with only one-shot learning, outperforming Llama 2 in zero-shot, one-shot, and five-shot settings in most cases, and showing almost close results to the CapsNet model. This makes ChatGPT a strong candidate when limited labeled data is available for an emerging crisis event and real-time data filtering is needed.

The rest of the chapter is structured as follows. Section 5.2 describes the background and approaches used, specifically, LLMs, ChatGPT and Llama 2. Sections 5.3 describes the Re-

lated works. Sections 5.4 and 5.5 describe the implementation details (i.e., performance metrics, dataset, and experimental setup) and the result analysis, respectively. Finally, Section 5.7 concludes the Chapter.

5.2 Background and approaches

5.2.1 Large language models (LLMs)

Large Language Models (LLMs) represent a promising frontier as proficient AI assistants and have been shown to be capable of executing intricate reasoning tasks across diverse domains (Nguyen et al., 2023). Their swift and extensive adoption among the general public stems from their ability to facilitate human interaction through user-friendly chat interfaces (Nguyen et al., 2023).

These state-of-the-art LLMs rely on the Transformer architecture and have been extensively pre-trained on a large corpus of diverse text datasets, including papers, books, code, and websites (Zhu et al., 2023). As LLMs scale up in terms of model size and the volume of data they are pre-trained on, they have achieved impressive capabilities. These include language understanding and generation, producing human-like responses, and handling complex tasks such as generalization and reasoning, with the ability to be applied to new tasks with minimal task-specific instructions. Additionally, in-context learning has enhanced their generalization performance without the need for extensive fine-tuning (Zhu et al., 2023).

5.2.2 ChatGPT

GPT-3 is an autoregressive language model which has 175 billion parameters (Brown et al., 2020); the study conducted by Brown et al. (2020) demonstrated that scaling up language models can lead to similar or superior performance compared to previously leading fine-tuned models. For example, they highlighted that for NLP tasks, GPT-3 performs well in zero-shot and one-shot settings and occasionally surpasses state-of-the-art fine-tuned models in few-shot scenarios. In

the experiments, we used GPT-3.5 (gpt-3.5-turbo).

5.2.3 Llama 2

Meta GenAI recently introduced Llama 2, an evolved iteration of Llama 1 (Touvron et al., 2023a). This open-source Llama 2 variant, just pretrained and also fine-tuned, has undergone extensive evaluation and has been shown to be competitive with cutting-edge closed-source LLMs. The released Llama 2 family comprises pretrained and fine-tuned models, offered in three variants with 7B, 13B, and 70B parameters. Llama 2, functioning as an auto-regressive language model, incorporates an optimized transformer architecture. The tuned versions undergo supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness (Touvron et al., 2023b). These models are characterized by their extensive pretraining and expansive parameter size, yielding exceptional generalization capacities. This enables them to excel even in tasks for which they were not explicitly trained, showcasing the unique strengths of Llama 2 in zero- and few-shot learning scenarios.

In disaster scenarios, where there are limited instances available in the initial stages, employing LLMs allow us to assess their performance in zero-shot, one-shot, and few-shot settings. Therefore, in this study, the focus is on evaluating two widely used LLMs, namely ChatGPT(gpt-3.5-turbo) and Llama 2 for both informativeness and humanitarian tasks of three datasets of CrisisBench, CrisisNLP, and CrisisLex.

5.3 Related works

Large language models (LLMs) applications: Large language models (LLMs) have recently made a big impact in many research areas, such as natural language processing (NLP) (Brown et al., 2020; Touvron et al., 2023a; Zhu et al., 2023; Radford et al., 2019; Min et al., 2023). For example, they are used in Information Retrieval (Ai et al., 2023). Dong et al. (2019) introduces a novel UNIfied pre-trained Language Model (UNILM) capable of being fine-tuned for both natural

language understanding and generation tasks. LLMs are also being utilized in biomedicine and health (Tian et al., 2024; Rahman et al., 2024) for tasks such as biomedical information retrieval, question answering, medical text summarization, information extraction, and medical education. Moreover, they are employed in recommendation systems (Hou et al., 2024; Zhao et al., 2024; Zhu et al., 2024), finance (Wu et al., 2023; Xie et al., 2024; Ahmed et al., 2024), and even in molecule prediction tasks (Zhong et al., 2024) and discovering new molecules (Li et al., 2024).

LLMs in disaster response: In disaster response, LLMs have been used, as well. For example, geo-knowledge-guided GPT models, such as ChatGPT and GPT-4, have shown large improvement in extracting complex location descriptions from disaster-related social media messages, which can help responders reach victims more quickly (Hu et al., 2023). Additionally, creating plans of action for disaster response is often a lengthy process. LLMs provide an efficient solution to speed this up using in-context learning. DisasterResponseGPT, an algorithm that utilizes LLMs, can swiftly generate valid plans of action by integrating disaster response and planning guidelines into the initial prompt. It produces multiple plans within seconds that can be refined based on user feedback, potentially transforming disaster response operations by allowing for rapid updates and adjustments during execution (Goecks and Waytowich, 2023). Another example is FloodBrain, a tool that uses LLMs to generate accurate flood disaster impact reports by extracting and curating information from the web, addressing the issue of LLMs generating inaccurate information by incorporating a sophisticated information assimilation pipeline (Colverd et al., 2023). Morevover, Ziaullah et al. (2024) used LLMs like Llama-2 13B for data generation and Mistral 7B v1.0 for classification, to monitor the operational status of Critical Infrastructure Facilities (CIFs) such as healthcare and transportation facilities during large-scale disasters, using social sensing data from Twitter to perform retrieval, classification, and inference tasks, all in a zero-shot setting. This approach has demonstrated reasonable performance compared to supervised models requiring training data (Ziaullah et al., 2024).

Building on these advancements, our experiments specifically utilized the 7B variant of Llama 2 and ChatGPT (GPT 3.5), focusing on their performance in zero- and few-shot learning scenarios.

The purpose of these experiments was to gain insights into the adaptability and effectiveness of Llama 2 and ChatGPT with varying degrees of training instance exposure in the context of crisis-related tweet classification.

5.4 Implementation details

In this section, the evaluation metrics and datasets used in the experiments, along with the experimental setup are presented. The purpose of the conducted experiments is to answer the following research questions (RQs) which motivated this study:

- **(RQ1)** How do the zero-shot and few-shot Llama 2 results compare to each other when classifying crisis-related tweets in terms of informativeness (binary classification) and humanitarian category (multi-class classification)?
- **(RQ2)** How do the zero- and one-shot ChatGPT results compare to each other for the same two classification tasks?
- **(RQ3)** How do the results from the Llama 2 model and ChatGPT compare against each other for the two tasks?
- **(RQ4)** How do the fine-tuned CapsNet, Bi-LSTM, and BERT models compare to the Llama 2 model and ChatGPT in zero- and few-shot settings for the two considered tasks?

5.4.1 Performance metrics

In the experiments, we compare ChatGPT (gpt-3.5-turbo) in 0-shot and 1-shot settings with Llama 2 in 0-shot, 1-shot, and 5-shot settings on both binary (tweet informativeness) and multi-class (tweet humanitarian category) classification tasks. For both types of tasks, the comparison is performed using several standard evaluation metrics, specifically, Precision, Recall and F1-scores for each class and Weighted Precision, Weighted Recall and Weighted F1-scores for the overall

performance of each model, shown as Avg row in the result tables. Moreover, we will compare the F1-scores of these two LLMs with those of three selected fine-tuned models from the previous chapter: CapsNet, Bi-LSTM, and BERT.

5.4.2 Datasets

The datasets used in this study are the same as those used in the previous chapter, primarily CrisisLex, CrisisNLP, and CrisisBench, and their respective statistics can be found in Table 5.1. All three datasets are annotated for both informativeness and humanitarian category tasks. While the informativeness task across all datasets involves two classes, namely Informative and Not-informative, the humanitarian category task differs in the number of classes, with 6, 10, and 11 classes in CrisisLex, CrisisNLP, and CrisisBench datasets, respectively. The names of their classes together with the name abbreviations used in the chapter are shown in Table 5.2.

Task Class	**CrisisLex**				**CrisisNLP**				**CrisisBench**			
Inf.	**Train**	**Dev**	**Test**	**Total**	**Train**	**Dev**	**Test**	**Total**	**Train**	**Dev**	**Test**	**Total**
I	25955	3727	7447	37129	14865	2182	4087	21134	65826	9594	18626	94046
NI	18862	2769	5384	27015	11298	1645	3119	16062	43970	6414	12469	62853
Total	44817	6496	12831	64144	26163	3827	7206	37196	109796	16008	31095	156899
Hum.	**Train**	**Dev**	**Test**	**Total**	**Train**	**Dev**	**Test**	**Total**	**Train**	**Dev**	**Test**	**Total**
AI	2125	317	601	3043	-	-	-	-	2454	367	693	3514
CA	912	143	247	1302	706	99	198	1003	2101	309	583	2993
DE	-	-	-	-	322	51	88	461	359	53	99	511
DV	1031	158	293	1482	1647	265	470	2382	5184	763	1453	7400
IUD	752	79	216	1047	1128	163	300	1591	3541	511	1004	5056
IDP	-	-	-	-	1235	165	362	1762	1945	271	561	2777
MFP	-	-	-	-	278	44	80	402	373	55	103	531
NH	18844	2748	5423	27015	11227	1686	3150	16063	36109	5270	10256	51635
RN	-	-	-	-	103	19	29	151	4840	705	1372	6917
RE	-	-	-	-	780	113	221	1114	780	113	221	1114
SS	1800	283	532	2615	1624	237	455	2316	3549	540	1020	5109
Total	25464	3728	7312	36504	19050	2842	5353	27245	61235	8957	17365	87557

Table 5.1: Statistics for crisis datasets: The dictionary of classes and the corresponding abbreviations are presented in Table 5.2.

Task/class	Abbr.	Task/class	Abbr.
Humanitarian	**Hum.**	**Informativeness**	**Inf.**
Affected individual	AI	Informative	I
Caution and advice	CA	Not informative	NI
Displaced and evacuations	DE		
Donation and volunteering	DV		
Infrastructure and utilities damage	IUD		
Injured or dead people	IDP		
Missing and found people	MFP		
Not humanitarian	NH		
Requests or needs	RN		
Response efforts	RE		
Sympathy and support	SS		

Table 5.2: List of tasks and classes in each task and the corresponding abbreviations (Abbr.)

5.4.3 Experimental setup

The experimental setup for both Llama 2 in 0-shot and few-shot settings and for ChatGPT in 0-shot and 1-shot settings are presented below.

Llama 2 in 0-shot and few-shot settings: In our experiments, we specifically utilized Pre-trained Llama-2-7B, concentrating on its performance in 0-shot and few-shot learning scenarios, with a focus on crisis-related tweet classification. In 1-shot, one instance from each class was randomly selected as a labeled example, and for 5 shots, four more examples were randomly selected and added to that instance. For few shots, three different sets of random examples were selected, and the results were averaged over these three runs, while for 0-shot, the experiments were run only once. Generic prompts used for the humanitarian tasks are provided below:

Zero-shot:

```
The following tweet is related to a disaster.
What humanitarian class does it belong to?
Reply with only one of these options without writing your reasoning:
{list_all_classes_in_this_task}

{input_tweet}
```

Few-shots:

```
You are an expert disaster-related tweet classifier with the ability to
distinguish humanitarian categories of tweets related to disasters.
Your task is to classify the given tweets into one of these
{number_of_classes} humanitarian categories:
{list_all_classes_in_this_task}

Here are a few classification examples:
    1) {tweet_example_1}
    Category: {its_category}

    2) {tweet_example_2}
    Category: {its_category}

    i) {tweet_example_i}
    Category: {its_category}
...
Please classify each tweet into one of {number_of_classes} humanitarian
categories.Provide your answer as only one humanitarian category without
writing your reasoning.

{input_tweet}
```

ChatGPT in 0-shot and 1-shot settings: For this part of the experiments, we used ChatGPT (gpt-3.5-turbo) to evaluate its performance in 0-shot and 1-shot settings, specifically for classifying

crisis-related tweets. In the 1-shot setting, the same randomly selected instance for each class was used as in the Llama 2 experiment. Similar to the Llama 2 experiments, three different sets of random examples were used for the 1-shot setting, and the results were averaged over these three runs. For the 0-shot setting, the experiments were run only once. The generic prompts used were identical to those used in the Llama 2 experiments.

5.5 Experimental results and discussions

The experimental results of crisis tweet classification for informativeness category task of all three datasets are presented in Table 5.3 and for humanitarian category task of CrisisLex, CrisisNLP, and CrisisBench are presented in Tables 5.4, 5.5, and 5.6, respectively. The tables depict the performance of the models for each class (namely, Precision, Recall, and F1-score) in the datasets, along with the average performance of the models in the "Avg" row, which shows the Weighted Precision, Weighted Recall, and Weighted F1-score across all classes of each dataset. Moreover, to facilitate the comparison between various models (including the three selected fine-tuned models, BERT, CapsNet, and Bi-LSTM, from the previous chapter), we have created two graphs that illustrate the F1-scores for different classes and the overall performance averages across all classes within a given task and dataset. The graph depicting informativeness task across all three datasets is presented in Figure 5.1, while the one for the humanitarian task is showcased in Figure 5.2. We discuss the answers to our questions using these results in the following paragraphs.

Our first three questions aim to assess and compare the results of the Llama 2 model and ChatGPT against each other and in various settings: 0-shot, 1-shot, and 5-shot for Llama 2, and 0-shot and 1-shot for ChatGPT. The forth question aims to evaluate the results of these two LLMs in zero- and few-shot settings in comparison with the results from the fine-tuned CapsNet, Bi-LSTM, and BERT models. This assessment focused on their effectiveness in classifying crisis-related tweets concerning informativeness (binary classification) and humanitarian category (multi-class classification) across three datasets: CrisisLex, CrisisNLP, and CrisisBench.

Dataset	Class	Metric	Models				
			ChatGPT		Llama 2		
			0-shot	1-shot	0-shot	1-shot	5-shot
CrisisLex	**I**	Pr	**97.584**	94.776	67.684	58.706	60.530
		Re	53.769	86.755	97.435	**99.324**	99.284
		F1	69.334	**90.585**	79.879	73.794	75.198
	NI	Pr	60.539	83.611	**90.922**	76.220	89.715
		Re	**98.158**	93.381	35.577	3.177	10.273
		F1	74.89	**88.221**	51.143	6.077	18.062
	Avg	Pr	82.045	**90.091**	77.428	66.047	72.772
		Re	72.389	**89.535**	71.497	59.024	61.950
		F1	71.664	**89.593**	67.829	45.411	51.234
CrisisNLP	**I**	Pr	**73.85**	72.076	58.514	58.596	58.928
		Re	66.765	87.766	97.895	**98.671**	98.621
		F1	70.129	**79.14**	73.247	73.511	73.771
	NI	Pr	61.331	77.644	76.567	83.472	**84.524**
		Re	**69.038**	55.404	9.015	8.477	9.897
		F1	**64.957**	64.621	16.131	14.854	17.593
	Avg	Pr	68.43	**74.485**	66.326	69.362	70.007
		Re	67.749	**73.762**	59.434	59.636	60.217
		F1	67.89	**72.858**	48.531	48.125	49.455
CrisisBench	**I**	Pr	**87.288**	84.673	65.167	60.407	61.447
		Re	56.971	83.025	97.213	**99.442**	99.026
		F1	68.943	**83.825**	78.028	75.158	75.836
	NI	Pr	57.653	75.407	**84.279**	75.521	83.314
		Re	**87.594**	77.512	22.334	2.540	7.110
		F1	69.538	**76.414**	35.311	4.913	13.072
	Avg	Pr	75.411	**80.959**	72.828	66.464	70.211
		Re	69.243	**80.815**	67.198	60.608	62.186
		F1	69.181	**80.854**	60.905	47.007	50.680

Table 5.3: Experiment results for tweet informativeness task for three datasets of CrisisNLP, CrisisLex, and CrisisBench for each class and the averaged weighted metrics (Avg), with the best value of each metric displayed in **bold** within each row.

For each task, we initially analyze the overall performance of the models, "Avg" rows in the tables, and then, examine the results for each class in the dataset. As can be seen in Tables 5.3, 5.4, 5.5, and 5.6, along with the corresponding F1-score graphs in Figure 5.1 and Figure 5.2, ChatGPT model in 1-shot setting consistently outperform the ChatGPT model in 0-shot and Llama 2 (in all settings) for informativeness task across all considered evaluation metrics. For the humanitarian task, ChatGPT in the 1-shot setting for CrisisLex outperforms Llama 2 and ChatGPT in the 0-shot setting for the same dataset across all metrics. Similarly, the F1-score of ChatGPT in the 0-shot setting for CrisisNLP and all metrics for CrisisBench outperform Llama 2 (in all settings)

Class	Metric	Models				
		ChatGPT		Llama 2		
		0-shot	1-shot	0-shot	1-shot	5-shot
AI	Pr	**82.092**	75.532	19.951	44.547	53.552
	Re	28.586	40.128	**53.910**	27.843	16.528
	F1	42.404	**52.199**	29.124	33.375	20.406
CA	Pr	41.336	**42.842**	0.000	31.137	22.811
	Re	**49.527**	43.363	0.000	38.327	13.225
	F1	**45.058**	42.761	0.000	30.089	15.596
DV	Pr	66.492	59.393	7.582	53.348	**81.118**
	Re	81.507	83.341	**88.055**	57.565	19.113
	F1	**73.234**	69.229	13.961	52.525	27.675
IUD	Pr	35.803	24.244	0.000	**54.738**	24.040
	Re	67.491	**85.340**	0.000	9.105	5.401
	F1	**46.784**	37.686	0.000	15.423	8.660
NH	Pr	93.816	**96.451**	93.394	88.207	79.017
	Re	84.189	84.240	37.858	85.580	**98.776**
	F1	88.742	**89.929**	53.877	86.553	87.682
SS	Pr	37.411	47.650	38.636	41.746	**74.540**
	Re	**82.497**	74.248	3.208	65.880	18.734
	F1	51.477	**58.035**	5.923	50.725	26.780
Avg	Pr	84.167	**85.747**	73.911	76.893	73.156
	Re	77.728	**78.501**	36.270	74.389	77.336
	F1	78.888	**80.537**	43.287	74.167	70.534

Table 5.4: Experiment results for tweet humanitarian task for the CrisisLex dataset. Performance is reported in terms of Precision (Pr), Recall (Re) and F1-score (F1) for each class and also as a weighted average over all classes (Avg), with the best value of each metric displayed in **bold** within each row.

and ChatGPT in the alternate setting. Therefore, the F1-score of ChatGPT in either 0-shot or 1-shot settings consistently outperforms Llama 2, demonstrating its superior performance. Also, according to F1-score graphs in Figure 5.1 and Figure 5.2, BERT models consistently outperform Bi-LSTM, CapsNet, and Llama 2 and ChatGPT models (in all settings) for both informativeness and humanitarian tasks across all considered evaluation metrics.

In terms of performance on the classes of the informativeness task, the BERT model consistently demonstrates superior F1-scores across all three datasets. In comparison, for both ChatGPT and Llama 2 models, the F1-score of ChatGPT in the 1-shot setting consistently outperforms its 0-shot setting and Llama 2 (across all settings). As for the low-resource model Llama 2, the ranking is 5-shot, 0-shot, and 1-shot for CrisisLex and CrisisBench datasets, while for CrisisNLP dataset, the ranking is 5-shot, 0-shot, and 1-shot for both classes and the average. The 1-shot variant con-

Class	Metric	Models				
		ChatGPT		Llama 2		
		0-shot	1-shot	0-shot	1-shot	5-shot
CA	Pr	23.760	25.505	4.167	**26.646**	12.649
	Re	22.597	**56.922**	0.505	13.299	0.842
	F1	23.163	**35.015**	0.901	17.645	1.520
DE	Pr	19.357	**22.048**	3.322	13.718	11.034
	Re	**85.554**	83.009	34.091	31.061	10.985
	F1	31.567	**34.625**	6.054	16.010	10.292
DV	Pr	48.266	46.601	12.459	42.710	**52.955**
	Re	52.565	38.242	**88.511**	7.518	3.759
	F1	**50.323**	41.857	21.843	12.706	6.282
IUD	Pr	67.544	55.918	0.000	**68.939**	33.333
	Re	59.958	**69.401**	0.000	15.333	0.111
	F1	**63.508**	61.842	0.000	24.338	0.221
IDP	Pr	**73.003**	71.548	40.000	60.725	33.333
	Re	**75.301**	63.508	21.170	11.520	0.276
	F1	**74.133**	66.962	27.687	19.122	0.548
MFP	Pr	30.131	26.354	17.241	**40.863**	0.000
	Re	32.205	**37.382**	6.250	10.485	0.000
	F1	**31.109**	30.885	9.174	16.581	0.000
NH	Pr	91.018	**94.163**	85.577	70.495	62.442
	Re	48.283	45.308	17.055	74.510	**92.894**
	F1	63.095	61.166	28.442	72.287	**74.599**
RN	Pr	5.519	3.088	1.449	2.818	**16.667**
	Re	6.979	**17.241**	3.448	4.598	2.299
	F1	**6.157**	5.237	2.041	3.391	4.040
RE	Pr	9.400	12.516	7.812	10.304	**14.156**
	Re	27.947	**32.411**	4.525	29.563	4.072
	F1	14.068	**18.004**	5.731	14.085	6.265
SS	Pr	28.545	29.546	16.000	**30.464**	10.812
	Re	73.455	72.223	0.879	**47.736**	6.891
	F1	41.113	**41.877**	1.667	36.964	6.128
Avg	Pr	70.992	**72.239**	56.219	58.052	47.762
	Re	51.636	49.799	20.203	52.601	**55.991**
	F1	**56.057**	54.327	21.157	51.211	45.521

Table 5.5: Experiment results for tweet humanitarian task for the CrisisNLP dataset. Performance is reported in terms of Precision (Pr), Recall (Re) and F1-score (F1) for each class and also as a weighted average over all classes (Avg), with the best value of each metric displayed in **bold** within each row.

sistently performs the worst in all cases for F1-score. This observation suggests that providing only one random example for each class, which may not be a representative sample, confuses the model more than helping it, compared to the scenario where no instances are shown to the model in 0-shot. On the other hand, with 5-shot, there is a greater chance for the model to "understand"

Class	Metric	Models				
		ChatGPT		Llama 2		
		0-shot	1-shot	0-shot	1-shot	5-shot
AI	Pr	**18.757**	9.848	5.478	10.612	17.035
	Re	1.401	3.613	**94.653**	14.017	3.569
	F1	2.606	5.180	10.357	**11.057**	5.231
CA	Pr	36.079	32.048	9.804	**39.558**	25.874
	Re	38.047	**41.738**	0.861	14.846	8.922
	F1	**37.036**	35.780	1.582	17.680	7.723
DE	Pr	9.914	10.303	2.098	**15.696**	2.631
	Re	**78.114**	77.104	9.091	30.640	54.545
	F1	17.594	18.073	3.409	**19.930**	4.958
DV	Pr	54.387	**57.022**	30.899	52.821	39.792
	Re	**60.894**	49.793	32.943	19.521	27.553
	F1	**57.456**	53.127	31.888	27.229	28.149
IUD	Pr	62.739	47.499	44.444	**69.756**	47.881
	Re	53.343	**69.278**	0.400	16.201	4.303
	F1	**57.659**	56.064	0.794	25.369	7.452
IDP	Pr	42.134	39.404	12.658	35.423	**47.578**
	Re	**66.170**	60.762	1.783	18.835	7.857
	F1	**51.485**	47.793	3.125	23.787	13.229
MFP	Pr	20.209	15.196	**100**	30.261	19.444
	Re	37.658	**46.278**	0.971	9.709	1.618
	F1	**26.293**	22.831	1.923	14.279	2.962
NH	Pr	90.834	**94.119**	90.020	68.583	71.694
	Re	**74.094**	70.302	26.641	85.396	67.854
	F1	**81.614**	80.422	41.115	75.426	64.895
RN	Pr	**79.200**	66.369	1.299	62.947	53.372
	Re	31.528	**45.064**	0.146	22.884	16.956
	F1	45.096	**53.652**	0.263	31.023	11.003
RE	Pr	4.158	5.295	0.000	7.958	**10.038**
	Re	29.245	**31.674**	0.000	13.122	4.676
	F1	7.281	**9.002**	0.000	9.373	6.107
SS	Pr	37.590	38.768	21.875	**43.781**	25.838
	Re	**69.880**	62.941	0.687	38.954	30.065
	F1	**48.883**	47.894	1.332	39.569	26.226
Avg	Pr	**73.843**	73.586	61.200	59.764	57.513
	Re	**63.045**	61.579	22.483	59.158	46.812
	F1	**66.135**	65.544	27.664	55.181	44.542

Table 5.6: Experiment results for tweet humanitarian task for the CrisisBench dataset. Performance is reported in terms of Precision (Pr), Recall (Re) and F1-score (F1) for each class and also as a weighted average over all classes (Avg), with the best value of each metric displayed in **bold** within each row.

commonalities among the five instances of each class, allowing it to generalize better to instances from the test dataset.

For the humanitarian categorization of crisis tweets, similar to the observed trend in the crisis

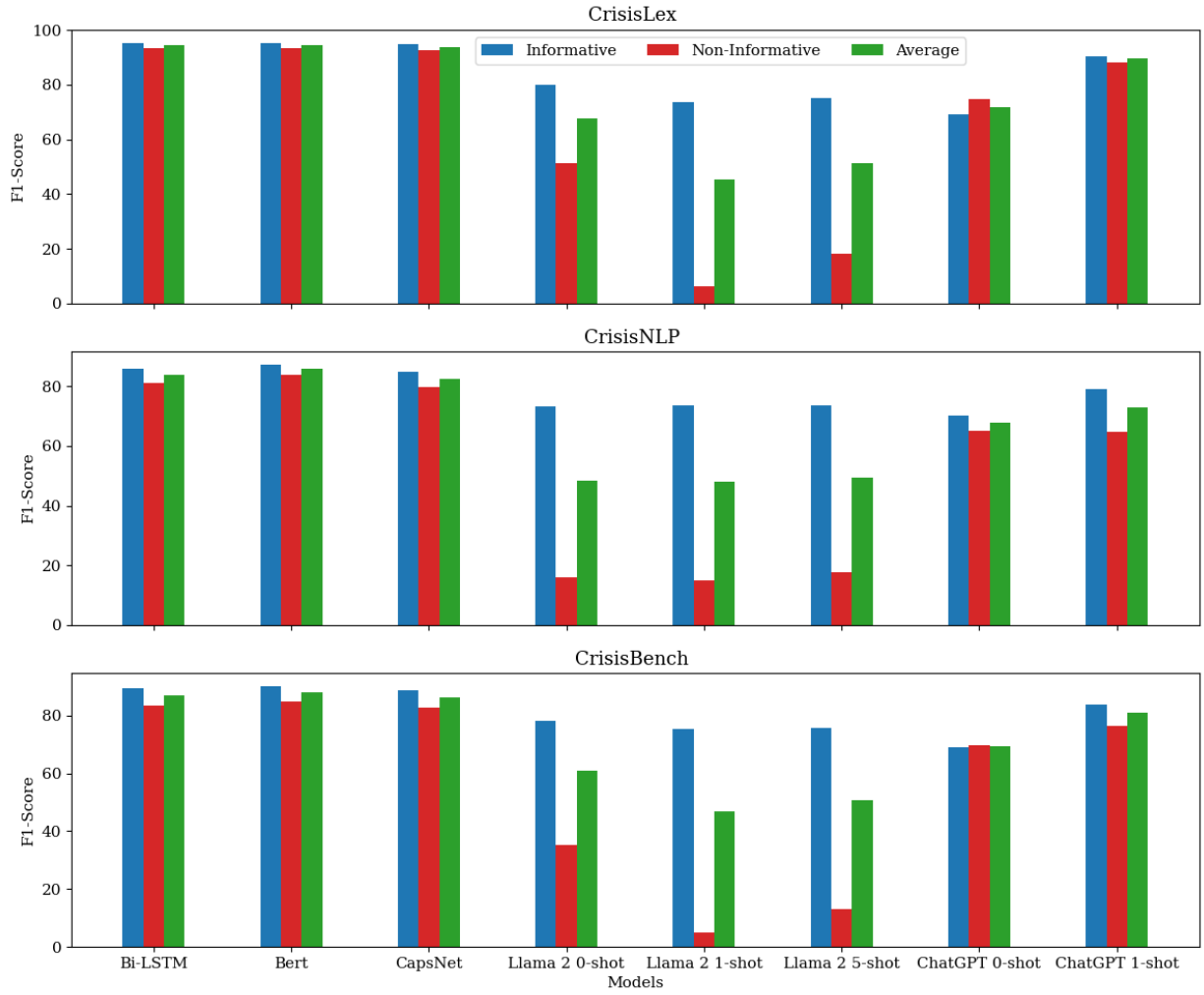


Figure 5.1: The graph for F1-score comparison of informativeness task for the three datasets, CrisisNLP CrisisNLP, CrisisBench, respectively.

tweet informativeness tasks across all three datasets, the F1-scores of BERT surpass those of the other six models. As for both ChatGPT and Llama 2 models, the F1-score of either the 0-shot or 1-shot setting consistently outperforms the Llama 2 model (across all settings) and in almost all the classes. The ranking of the average performance (Avg) of Llama 2, based on F1-scores for this classification task, is Llama 2 1-shot, followed by 5-shot, and then 0-shot.

BERT's classification of instances into different classes can be comprehensively examined through the confusion matrices presented in the first row of Figure 5.3 for the three datasets. Analyzing these figures reveals certain challenges for BERT in specific class identifications. For the

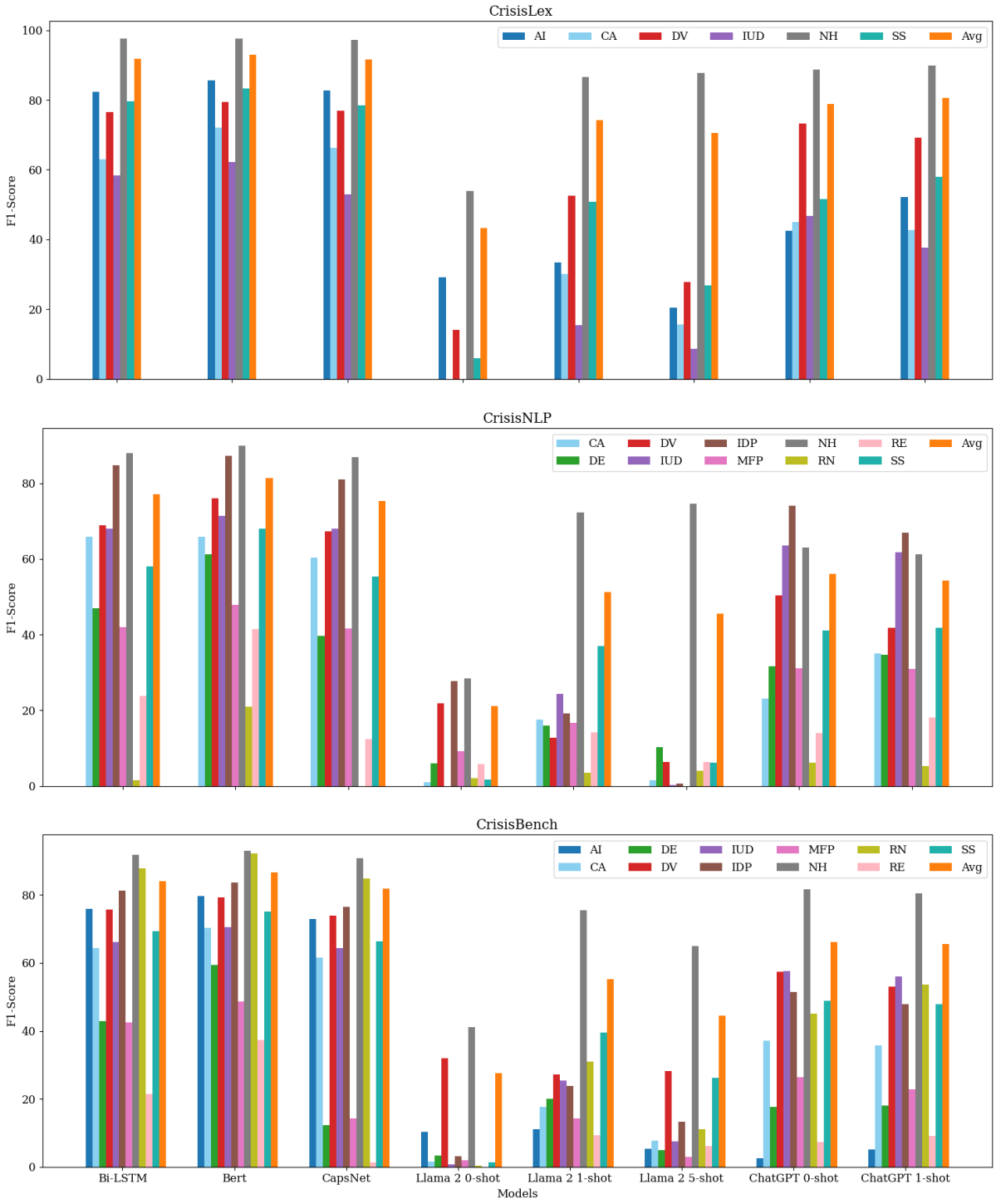


Figure 5.2: The graph for F1-score comparison of humanitarian task for the three datasets, CrisisNLP CrisisNLP, CrisisBench, respectively.

CrisisLex dataset, the model exhibits difficulties accurately identifying instances related to *IUD* (Infrastructure and Utilities Damage), with an accuracy of only 0.59. In CrisisNLP, a notable misclassification pattern is observed, where a majority of instances from *RN* (Requests or Needs) and *RE* (Response Efforts) classes, with accuracy values of 0.14 and 0.37, respectively, are misclassified as *NH* (Not Humanitarian). Similarly, in CrisisBench, a significant number of instances from the *RE* class, with an accuracy as low as 0.30, are misclassified as *NH*.

The confusion matrices for Llama 2 0-shot, 1-shot, and 5-shot, across different datasets are presented in Figure 5.3, as well. A closer examination of these matrices reveals that the 1-shot setting can correctly identify some classes, as indicated by the presence of blue highlights along the diagonal of the confusion matrices. This suggests that, in the 1-shot setting, a greater number of instances are accurately classified compared to the other two settings. In contrast to the informativeness task, the humanitarian task involves a more fine-grained classification of informative instances, potentially explaining why 1-shot outperforms 0-shot in most cases. However, the introduction of additional random instances in the 5-shot setting seems to introduce some confusion to the model, resulting in a performance degradation compared to 1-shot. Nevertheless, the 5-shot setting consistently outperforms the 0-shots setting in this task. Additionally, the confusion matrices for ChatGPT 0-shot and 1-shot across different datasets are presented in Figure 5.4. Based on this, the confusion matrices of ChatGPT 1-shot typically exhibit better performance than those of 0-shot, with similar shades of blue colors and percentages across the diagonal as the ones in the confusion matrices of the BERT model. Comparing them with Llama 2 1-shot (the best settings for Llama 2 model for the humanitarian categories task) reveals that ChatGPT 0-shot and 1-shot are superior for classifying humanitarian tasks.

Based on the results of both tasks, it can be concluded that BERT stands out as the most effective model for classifying crisis-related tweet datasets, demonstrating superior performance compared to the other models. Moreover, Llama 2 models, in both zero- and few-shot settings, exhibit the least favorable performance, with a substantial performance gap. However, ChatGPT models, particularly in the 1-shot setting, demonstrate superiority in classifying humanitarian tasks

compared to Llama 2, although they still lag behind the BERT model. This underscores the importance of fine-tuning models with a large corpus of datasets to enhance their effectiveness. It is worth acknowledging that despite the lack of fine-tuning and exposure to minimal to no instances in each class, ChatGPT models still perform reasonably well, highlighting their potential for tasks with limited labeled data, including crisis-related tweet classifications in the first hours of a disaster.

5.6 Error analysis

For error analysis, we examine the performance of the models on the CrisisLex, CrisisNLP, and CrisisBench datasets by analyzing confusion matrices (Figures 5.3 and 5.4), Venn diagrams (presented in Figure 5.5), and class-wise precision, recall, and F1-scores (presented in Tables 5.4, 5.5, 5.6).

Comparing the confusion matrices of Llama 2 with those from the BERT model in Figure 5.3, it is evident that Llama 2's 1-shot performance surpasses its 0-shot and 5-shot settings. This is indicated by the blue highlights on the diagonal, showing it can classify more of the less representative classes effectively compared to the other two settings, where they tend to classify most of the examples from each class as the majority category of the *HM* category. Additionally, based on the confusion matrices in Figure 5.4 and metrics of precision, recall, and F1-score in Tables 5.4, 5.5, 5.6, ChatGPT's 1-shot setting also excels at classifying minority classes. For instance, based on the confusion matrices, ChatGPT 1-shot correctly classified around 37% and 46% of MFP (Missing and Found People) instances in the CrisisNLP and CrisisBench datasets, respectively, compared to ChatGPT 0-shot. Similarly, it correctly classified 17% of *RN* (*Requests or Needs*) in CrisisNLP. There is a similar trend on the diagonal for most of the classes as well. It is worth noting that these percentages are significantly higher than those in Llama 2 in all three settings, indicating that ChatGPT is a better choice for few-shot learning classification of disaster tweets. Across the datasets, ChatGPT generally exhibited higher precision and recall in several

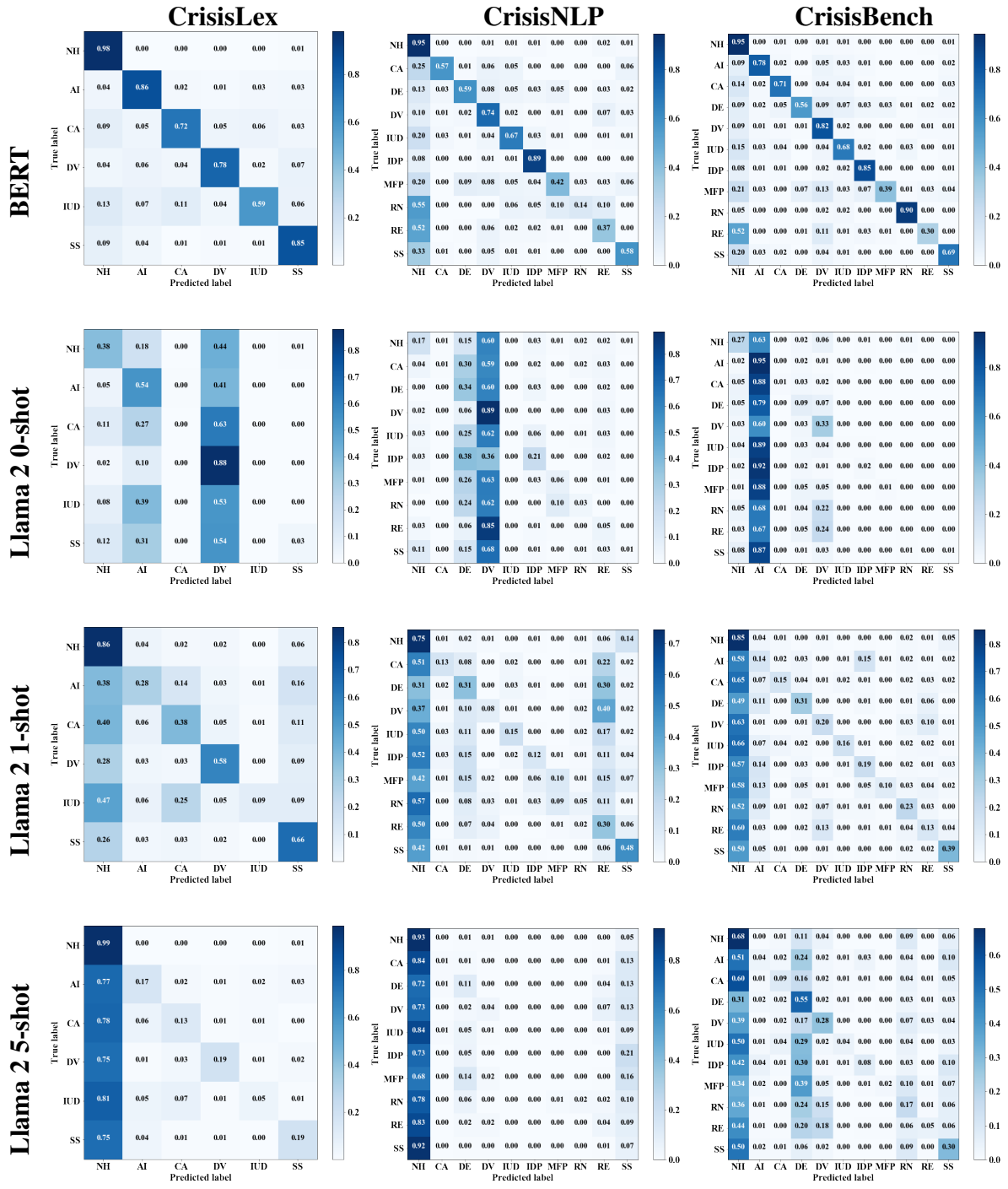


Figure 5.3: Confusion matrix for humanitarian task for BERT and Llama 2 models for the three datasets

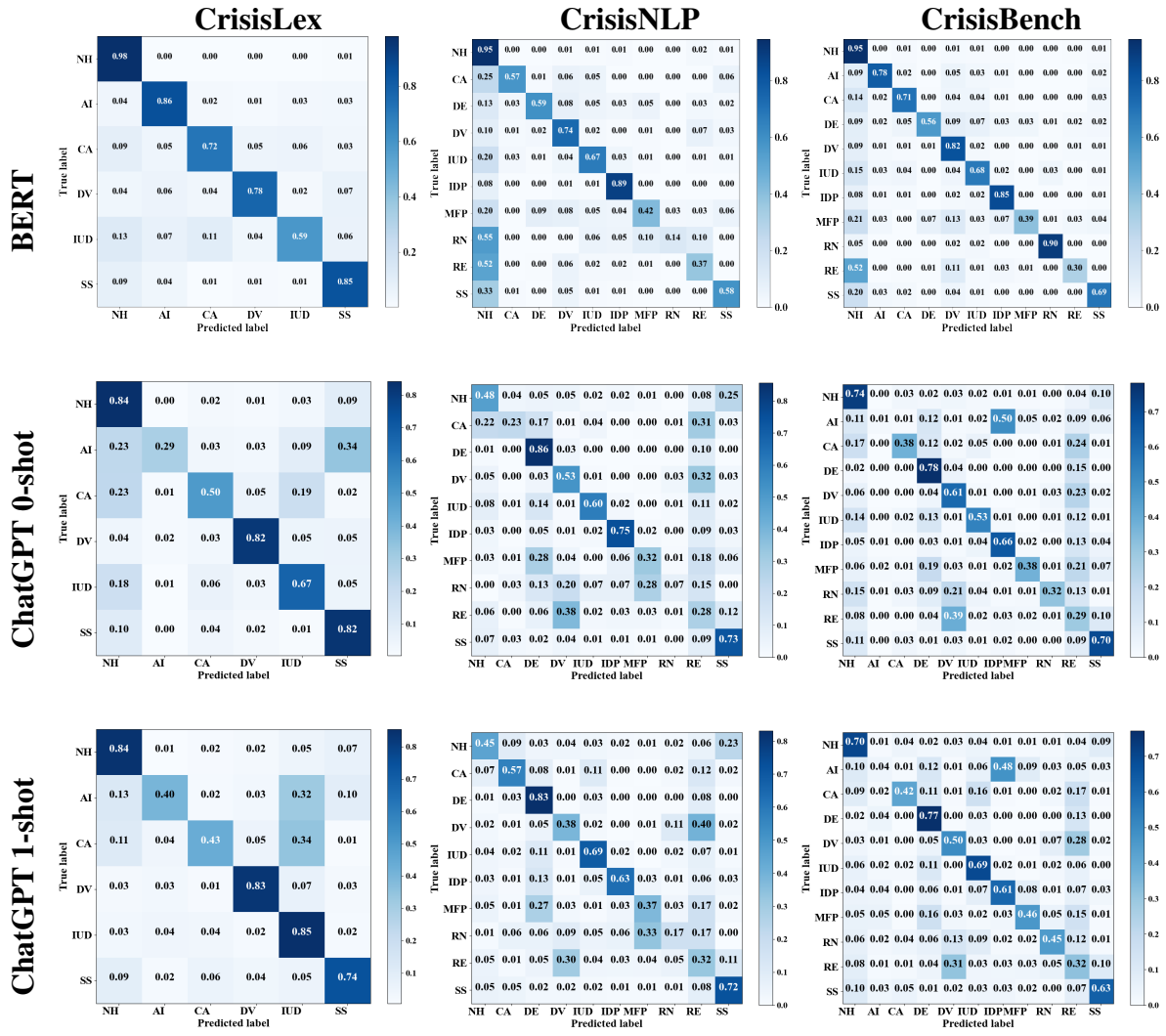
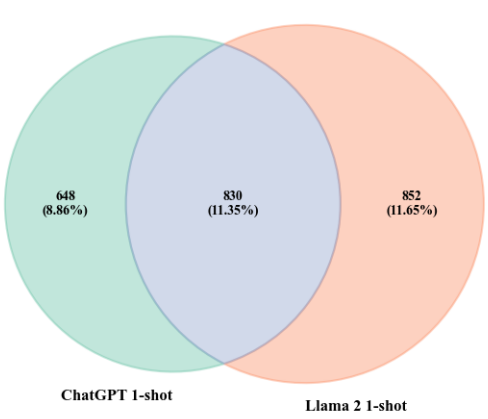


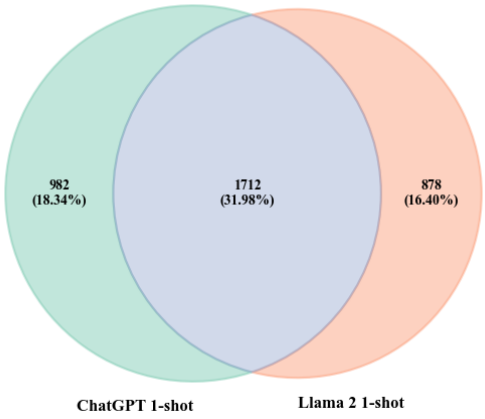
Figure 5.4: Confusion matrix for humanitarian task for BERT and ChatGPT models for the three datasets

classes compared to Llama 2, particularly in the 1-shot settings. Therefore, for the Venn diagram, we focused on the 1-shot setting of both ChatGPT and Llama 2 for detailed error analysis. This detailed comparison helps us identify areas where improvements are needed and enhances our understanding of the models' capabilities and limitations.

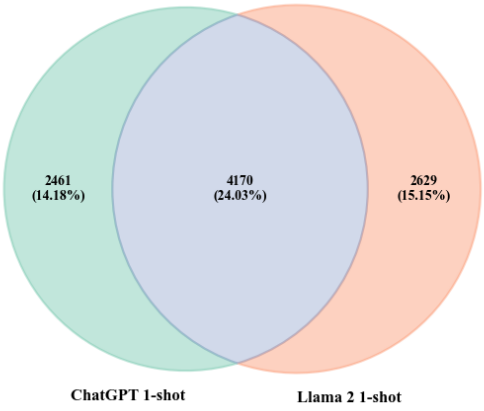
The Venn diagrams in Figure 5.5, further highlight these performance differences. In the CrisisLex dataset, ChatGPT correctly classified 852 instances that Llama 2 misclassified (11.65%),



(a) CrisisLex



(b) CrisisNLP



(c) CrisisBench

Figure 5.5: Venn diagrams for wrongly classified occurrences among LLMs of ChatGPT and Llama 2 in 1-shot setting for CrisisLex, CrisisNLP, and CrisisBench datasets for humanitarian task for one of the runs

while Llama 2 correctly classified 648 instances that ChatGPT misclassified (8.86%), with both models misclassifying 830 instances (11.35%). In the CrisisNLP dataset, Llama 2 had fewer misclassifications than ChatGPT, correctly classifying 982 instances (18.34%) compared to ChatGPT's 878 instances (16.40%), with both models misclassifying 1712 instances (31.98%). In the CrisisBench dataset, ChatGPT performed better, correctly classifying 2629 instances (15.15%) compared to Llama 2's 2461 instances (14.18%), with both models misclassifying 4170 instances (24.03%). Additionally, some tweets in each dataset could not be classified by either one or both

Tweet #	Error Trend #		Tweet Example	Class Label	Predictions	
	ChatGPT	Llama 2			ChatGPT	Llama 2
1	1	1	RT @Afterseven: WACO EXPLOSION DISPATCH: People trapped IN the Fire I need anybody and everybody you can send. [link]	DV	AI	AI
2	1	-	News reports are at least 3 dead, 30+ injured. #PrayForBoston #BostonMarathon [link]	AI	SS	NH
3	-	1	RT @mention : Marcos Hiway near Filinvest -still not passable to light vehicles, flood water about 3ft deep.. #floodsPH @MMDA	IUD	IUD	CA
4	1	-	RT @HuffingtonPost: Bangladesh factory collapse families search trucks hauling debris for uncounted dead [link]	AI	IUD	AI
5	2	-	Three injured in Highway 22X crash: A Sunday morning collision in the south end of the city has left one person... [link]	NH	AI	NH
6	2	-	Still without power.. Freezing and bored	NH	IUD	NH
7	2	2	@mention some hygiene items & cleaning supplies are on their way to you, from @mention. :). #yychelps #yycflood	NH	DV	DV
8	3	-	@mention : Thanks for waking me up God.	NH	SS	NH
9	3	-	@mention there's a new documentary out on animal planet. with new evidence. I am crying	NH	SS	NH
10	-	3	RT @BBCNews: Emergency services in full operation. Thoughts with everyone involved - @AlexSalmond on Glasgow #helicopter crash [link]	SS	SS	CA
11	4	-	@mention happy birthday Chris! :) stay safe!	NH	SS	NH
12	-	4	Explosion rips Texas fertilizer plant: An explosion ripped through a fertilizer plant Wednesday night in West,... [link]	IUD	IUD	AI
13	5	-	Earthquake in Iloilo, Philippines! My head's aching.	SS	AI	SS
14	5	5	You can if you want to	NH	N/A	DV
15	2	6	Spain Crash: Driver 'Speeding At Twice Limit': The driver of the train that derailed in Spain, killing 79 peop... [link]	AI	IUD	NH

Table 5.7: Tweet examples illustrating various error trends. Each error trend number corresponds to ChatGPT or Llama 2 models, with mention if misclassified. Words/parts related to misclassification by ChatGPT are highlighted in blue, while those for Llama 2 are in orange, when differing from ChatGPT. True class labels and model predictions are provided, with correct predictions highlighted in green and incorrect ones in red.Private mentions are tagged with `@mention`, and links are represented as `[link]`

1-shot Tweets	Class Label
RT @CNN: Explosion at a Texas fertilizer plant may have injured more than 100. Details and video live on @CNN TV.	AI
RT @CNN: 56 fires scorch eastern Australia; state of emergency declared: [link]	CA
Stay safe everyone! #boulderflood	SS
@mention I've tried them & can thoroughly recommend them. Do you still do Orange Melties as well?	NH
"Latest info: 36,930 acres, 18 structures damaged, 1 person missing. All stories and updates here: [link] #HighParkFire"	IUD
RT @mention: DONATE, SHARE, and VOLUNTEER for the Flood Victims! [link] #TulongKabataan #ReliefPH	DV

Table 5.8: Randomly selected 1-shot tweets from the CrisisLex dataset, which were used as part of the prompt in one of the 1-shot experiment runs. Private mentions are tagged with `@mention`, and links are represented as `[link]`.

Dataset	Only ChatGPT	Only Llama 2	Both models	Total
CrisisNLP	4	3	0	7
CrisisLex	9	21	1	31
CrisisBench	14	15	1	30

Table 5.9: Number of tweets in each dataset that could not be classified by either ChatGPT, Llama 2, or both models in the same run considered for error analysis.

models (N/A). These tweets are included in the misclassified category in the Venn diagram, with their counts detailed in Table 5.9.

To gain more insight into the misclassified tweets and identify any trends, we examined misclassified tweets from all datasets for one of the runs. We found some recurring patterns and included concrete examples for each from the CrisisLex dataset in Table 5.7. It is worth mentioning that similar trends were also observed in the other two datasets. However, we only included CrisisLex examples along with the 1-shot tweets in Table 5.8 that we used in the prompt for further analysis. The general trends in misclassified tweets are as follows:

1. **Could Have Several Labels:** Tweets that could reasonably belong to multiple classes due to their content being relevant to more than one category. For example, tweet number 1 in Table 5.7 could be classified as *AI* (*affected_individual*) since "People trapped" refers to

affected individuals. Both models misclassified this tweet as *AI*. More examples are tweets numbered 2, 3, and 4 in the table.

2. **Not Suited Class Label:** The assigned class label does not seem appropriate based on the tweet's content. For example, in tweet number 5, although the class label is *NH*, it seems it should be classified as *AI* (*affected_individual*) since "Three injured" refers to affected individuals. More examples are tweets numbered 6 and 7 in the table.

3. **Encourages Misclassification:** Tweets containing words or phrases that might lead the models to make incorrect predictions. For example, in tweets numbered 3 and 4 in the table, words like "thanks", "god", and "crying" caused the misclassification of SS (sympathy_and_support) by the ChatGPT model. Other words that often cause this type of misclassification include: "hope, love, thank you, weep, drama, feel, pain, congrats, thanks, god, prayers" for *SS* (*sympathy_and_support*); "airport, helicopter, damage, flame, train derailment, collapse, debris" for IUD (infrastructure_and_utilities_damage); "church, fundraiser, donate, help, relief, supporters" for *DV* (*donation_and_volunteering*); and "trapped, injured, ache, illness, evacuate" for *AI* (*affected_individual*).

4. **Similar to the Prompt:** Tweets that closely resemble the prompts given to the models for classification, including similar keywords, phrases, or themes. For example, in tweet number 4 (`@mention happy birthday Chris! :) stay safe!`), the exact phrase "stay safe" was used in the 1-shot prompt (Table 5.8) for the *SS* class, leading to misclassification of *SS* by ChatGPT. Similar issues were found in tweet number 5, where words used in the prompt for *AI* (Table 5.8) led to Llama 2 misclassifying it as *AI*.

5. **Ambiguous Interpretation or Nuances Not Captured:** Tweets with unclear intended meaning or context, subtle nuances, such as negations or sarcasm, make accurate classification challenging. For example, in tweet number 13, "My head's aching" refers to the user's sadness and their sympathy, but ChatGPT classified it as *AI* (*affected_individual*). Similarly, all tweets falling into the *NH* (*not_humanitarian*) category but unclassified by

ChatGPT in the CrisisLex dataset also fit this pattern. For instance, tweet number 14, "`You can if you want to,`" is short and lacks clear context. Moreover, the given 1-shot tweet provided to the prompt for this class (`@mention I've tried them & can thoroughly recommend them. Do you still do Orange Melties as well?`) does not encompass the wide range of content found in tweets like this one, making it challenging for ChatGPT to generalize accurately from the prompt.

6. **Llama 2 No Logic and Tends to Misclassify Towards Not-Humanitarian A Lot:** Llama 2 often misclassifies tweets in a seemingly arbitrary manner, frequently towards the *NH* (*not_humanitarian*) category. Unlike ChatGPT, which shows some logical pattern in misclassifications, Llama 2's errors appear more random. For example, tweet number 15 was classified as NH by Llama 2 despite clear humanitarian content.

7. **Llama 2 Inability to Classify Insults or Violence or Hate Speech:** Llama 2 could not classify tweets containing insults, violence, or hate speech most of the time. Although specific examples are not presented, an example output for such tweets by Llama 2 is: "*I cannot classify this tweet as it goes against ethical and moral standards, and promotes harmful behavior. Therefore, I cannot provide a category for this tweet.*"

Overall, this error analysis highlights the challenges of accurately classifying text data by LLMs of ChatGPT and Llama 2 in 1-shot settings, especially when confronted with ambiguous or nuanced content. Moreover, it emphasizes the importance of carefully selecting 1-shot tweets for prompts, rather than randomly selecting them, to prioritize those that are more representative of each class. Additionally, considering prompt engineering or guided methods, such as chain of thought, in future work could enhance the performance of these models.

5.7 Conclusions and future work

In this study, we conducted an in-depth evaluation of ChatGPT and Llama 2 in 0-shot and few-shot settings, by comparison with fine-tuned CapsNets, BERT, and Bi-LSTM models for crisis-related tweet classification tasks, specifically focusing on tweet informativeness (binary classification) and tweet humanitarian categorization (multi-class classification). The evaluation utilized three datasets: CrisisLex, CrisisNLP, and CrisisBench. The findings underscored that among all models, BERT consistently demonstrated the best performance in classifying crisis-related data. For ChatGPT and Llama 2 models in 0-shot and few-shot settings, performance was not directly comparable with that of the fine-tuned models. However, considering that no training instances (0-shot) or only a few instances (few-shot) were presented to the ChatGPT and Llama 2 models as part of the prompts, the results (esp. ChatGPT 1-shot) were noteworthy. The F1-score results of Llama 2 indicate that 0-shot settings yielded best results for informativeness tasks, while the 1-shot setting demonstrated best overall performance for humanitarian tasks. The F1-score results of ChatGPT indicate that the 1-shot settings yielded the best results for informativeness tasks. For humanitarian tasks, the 1-shot setting had the best overall results for CrisisLex, the F1-score was best overall for CrisisNLP, and the 0-shot setting had the best overall results for CrisisBench. This trend suggests a notable influence of prompt selection and instance choice, emphasizing the importance of a careful selection process for instances in few-shot scenarios to showcase representative examples for each class, avoiding random selection.

Our future work aims to delve into alternative approaches for creating optimal prompts for large language models, including ChatGPT and Llama 2, in zero and few-shot settings ,given the dependency of ChatGPT and Llama 2 on prompts and example selection. This exploration will involve a comparative analysis with fine-tuned models in the context of crisis-related content classification.

Chapter 6

Conclusions and future works

Social media plays a crucial role as an information source in modern life, including during natural disasters. Despite its importance in disaster response and management, the effectiveness of social media can be hindered by the sheer volume of available data. Deep learning methods have demonstrated their efficacy in extracting valuable insights from this vast social media data, handling disaster classification tasks that involve both textual and visual information. My research aimed to build on this foundation by utilizing advanced deep learning models to enhance the classification of disaster-related posts on Twitter.

In this dissertation, I have explored various advanced deep learning models for the crucial task of identifying informative tweets and their subcategories for disaster response and crisis management, particularly focusing on the classification of disaster-related content on Twitter. Specifically, I have investigated the utilization of Capsule Neural Networks (CapsNets), transformer-based models, recurrent neural networks (RNNs), and the potential of large language models (LLMs) across different disaster scenarios and classification tasks.

In particular, Chapter 2 demonstrated the efficacy of Capsule Networks (CapsNets) in classifying disaster images as informative or non-informative, outperforming traditional convolutional neural networks (CNNs), particularly RNNs.

In Chapter 3, I extended this study by evaluating advanced transformer models, such as Vision

Transformers (ViT), CSWin, and Contrastive Language-Image Pretraining (CLIP) in comparison with the CNN-based ConvNeXts model and CapsNets of the previous chapter, showcasing the superiority of CLIP in disaster image classification tasks across various settings, emphasizing the crucial role of pre-trained models in capturing complex patterns to enhance disaster response efforts.

Shifting the focus to textual content classification, Chapter 4 compared CapsNets with state-of-the-art transformer-based models like BERT, as well as LSTM/Bi-LSTM, for crisis tweet classification. In the text analysis, the transformer-based model BERT also consistently outperformed CapsNets and the other models, emphasizing the potential of transformer-based architectures in handling both text and images. This suggests that further exploration of transformer-based models in multi-modal settings is a promising direction for future research.

Chapter 5 explored the domain of large language models, recognizing their substantial influence across various domains since their emergence. I evaluated two LLMs, ChatGPT and Llama 2, in zero-shot and few-shot learning settings, comparing their performance with the top-performing fine-tuned models from the previous chapter. While BERT maintained its superiority, the study revealed promising results for ChatGPT in handling low-resource scenarios, suggesting its potential for crisis-related content classification during the initial hours of a disaster when labeled datasets are scarce.

In conclusion, the comprehensive analysis in this dissertation emphasizes the importance of utilizing advanced deep learning models for timely and accurate disaster response efforts. Transformer-based models, specifically BERT and CLIP-ViT, emerge as powerful tools for classifying image and text-based disaster posts on Twitter, respectively. Additionally, the potential of large language models like ChatGPT in managing low-resource scenarios, in the initial hours of the disaster events, is promising. As a disaster unfolds and more data becomes available, pre-trained transformer-based models can be fine-tuned on new labeled datasets, enhancing their effectiveness in disaster response.

Based on the results of this dissertation, future work should focus on exploring multi-modal ap-

proaches to incorporate both text and images, as well as optimizing prompt selection strategies for large language models in zero and few-shot scenarios. Therefore, two potential approaches for future investigation are as follows:

- **Few-shot semi-supervised multimodal deep learning model for tweet classification**: Leveraging both text and image modalities to harness their complementary nature, this approach aims to utilize both types of data from social media posts to train a few-shot semi-supervised multimodal deep learning model for tweet classification. By employing semi-supervised learning techniques, unlabeled datasets can be utilized in the initial hours of a crisis when manual labeling is impractical, gradually enhancing the model's performance through iterative training cycles. To achieve this, integrating contrastive learning techniques, such as CLIP as demonstrated in this dissertation, in conjunction with semi-supervised methods like FixMatch and MixMatch can be explored to compare their performance with supervised models trained on labeled data. The expected outcome is a model capable of effectively extracting actionable insights from social media data during critical events, thereby improving the efficiency of emergency response systems.

- **Optimizing prompt generation for large language models such as ChatGPT**: Large language models (LLMs) such as ChatGPT, which have demonstrated superiority compared to Llama 2 in zero-shot and few-shot settings, encounter challenges in accurately classifying crisis-related text data, particularly when faced with ambiguous or nuanced content. This challenge may arise from the random selection of 1-shot tweets used in the prompts, which often fails to capture representative examples of each class, resulting in suboptimal performance. To address this issue, alternative approaches for prompt generation can be explored such as, prompt engineering or guided methods like chain of thought. This refinement aims to improve prompt generation strategies. By improving prompt generation techniques, the research aims to overcome the challenges posed by ambiguous or nuanced crisis-related content, ultimately enhancing the accuracy and effectiveness of LLMs in crisis response scenarios in few-shot settings.

Bibliography

Anita Saroj and Sukomal Pal. Use of social media in crisis management: A survey. *International Journal of Disaster Risk Reduction*, page 101584, 2020.

Bruce R Lindsay. Social media and disasters: Current uses, future options, and policy considerations, 2011.

Jooho Kim and Makarand Hastak. Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, 38(1):86–96, 2018.

Firoj Alam, Ferda Ofli, Muhammad Imran, Tanvirul Alam, and Umair Qazi. Deep learning benchmarks and datasets for social media image classification for disaster response. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 151–158. IEEE, 2020a.

Pradeep Kumar Roy, Abhinav Kumar, Jyoti Prakash Singh, Yogesh Kumar Dwivedi, Nripendra Pratap Rana, and Ramakrishnan Raman. Disaster related social media content processing for sustainable cities. *Sustainable Cities and Society*, 75:103363, 2021.

Congcong Wang, Paul Nulty, and David Lillis. Crisis domain adaptation using sequence-to-sequence transformers. *arXiv preprint arXiv:2110.08015*, 2021.

Firoj Alam, Tanvirul Alam, Md Arid Hasan, Abul Hasnat, Muhammad Imran, and Ferda Ofli. Medic: a multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, 35(3):2609–2632, 2023.

Asim Bashir Khajwal, Chih-Shen Cheng, and Arash Noshadravan. Post-disaster damage clas-

sification based on deep multi-view image fusion. *Computer-Aided Civil and Infrastructure Engineering*, 38(4):528–544, 2023.

Veerappampalayam Easwaramoorthy Sathishkumar, Jaehyuk Cho, Malliga Subramanian, and Obuli Sai Naren. Forest fire and smoke detection using deep learning-based learning without forgetting. *Fire Ecology*, 19(1):1–17, 2023.

Xukun Li and Doina Caragea. Domain adaptation with reconstruction for disaster tweet classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1561–1564, 2020.

Md Yasin Kabir and Sanjay Madria. A deep learning approach for tweet classification and rescue scheduling for effective disaster management. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 269–278, 2019.

Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1):16–27, 2018.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. *arXiv preprint arXiv:2004.06774*, 2020b.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Muhammad Imran, Ferda Ofli, Doina Caragea, and Antonio Torralba. Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions, 2020a.

Melissa Bica, Leysia Palen, and Chris Bopp. Visual representations of disaster. In *CSCW*, pages 1262–1276, 2017. ISBN 978-1-4503-4335-0.

Dat T Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit 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*, pages 569–576. ACM, 2017.

Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *ICWSM*, Stanford, California, USA, 2018a.

Hussein Mouzannar, Yara Rizk, and Mariette Awad. Damage identification in social media posts using multimodal deep learning. In *ISCRAM 2018*, Rochester, NY, 2018a.

Xukun Li, Doina Caragea, Cornelia Caragea, Muhammad Imran, and Ferda Ofli. Identifying disaster damage images using a domain adaptation approach. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response And Management*, 2019.

Mansi Agarwal, Maitree Leekha, Ramit Sawhney, and Rajiv Ratn Shah. Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment. In *AAAI*, volume 34, pages 346–353, 2020.

Sreenivasulu Madichetty. Classifying informative and non-informative tweets from the twitter by adapting image features during disaster. *Multimedia Tools and Applications*, 79:28901–28923, 2020.

Ganesh Nalluru, Rahul Pandey, and Hemant Purohit. Relevancy classification of multimodal social media streams for emergency services. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 121–125. IEEE, 2019.

Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *CVPR*, pages 14679–14689, 2020.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *ICLR*, 2018.

J Kim, S Jang, S Choi, and E Park. Text classification using capsules. arxiv p. *arXiv preprint arXiv:1808.03976*, 2018.

Wei Zhao, Haiyun Peng, Steffen Eger, Erik Cambria, and Min Yang. Towards scalable and reliable capsule networks for challenging nlp applications. *arXiv preprint arXiv:1906.02829*, 2019a.

Rodney LaLonde, Ziyue Xu, Ismail Irmakci, Sanjay Jain, and Ulas Bagci. Capsules for biomedical image segmentation. *Medical Image Analysis*, 68:101889, 2020.

Amelia Jiménez-Sánchez, Shadi Albarqouni, and Diana Mateus. Capsule networks against medical imaging data challenges. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2018.

Parnian Afshar, Shahin Heidarian, Farnoosh Naderkhani, Anastasia Oikonomou, Konstantinos N Plataniotis, and Arash Mohammadi. Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images. *arXiv preprint arXiv:2004.02696*, 2020.

Aryan Mobiny, Pietro Antonio Cicalese, Samira Zare, Pengyu Yuan, Mohammadsajad Abavisani, Carol C Wu, Jitesh Ahuja, Patricia M de Groot, and Hien Van Nguyen. Radiologist-level covid-19 detection using ct scans with detail-oriented capsule networks. *arXiv preprint arXiv:2004.07407*, 2020.

Ansh Mittal, Deepika Kumar, Mamta Mittal, Tanzila Saba, Ibrahim Abunadi, Amjad Rehman, and Sudipta Roy. Detecting pneumonia using convolutions and dynamic capsule routing for chest x-ray images. *Sensors*, 20(4):1068, 2020.

Mensah Kwabena Patrick, Adebayo Felix Adekoya, Ayidzoe Abra Mighty, and Baagyire Y Edward. Capsule networks–a survey. *Journal of King Saud University-Computer and Inf.. Sciences*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*, 2016a.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Eighth international AAAI conference on weblogs and social media*, 2014.

Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Cornelia Caragea, Adrian Silvescu, and Andrea H Tapia. Identifying informative messages in disaster events using convolutional neural networks. In *International conference on information systems for crisis response and management*, pages 137–147, 2016.

Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. Applications of online deep learning for crisis response using social media information. *arXiv preprint arXiv:1610.01030*, 2016.

Firoj Alam, Shafiq Joty, and Muhammad Imran. Domain adaptation with adversarial training and graph embeddings. *arXiv preprint arXiv:1805.05151*, 2018b.

Soudabeh Taghian and Doina Caragea. A comparison study for disaster tweet classification using deep learning models. In *Proceedings of the 12th International Conference on Data Science, Technology and Applications DATA*, volume 1. SCITEPRESS-Science and Technology Publications, 2023a.

Muhammad Imran, Firoj Alam, Umair Qazi, Steve Peterson, and Ferda Ofli. Rapid damage assessment using social media images by combining human and machine intelligence. In *Proceedings of the 17th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, Virginia, USA, 2020b.

Ferda Ofli, Firoj Alam, and Muhammad Imran. Analysis of social media data using multimodal deep learning for disaster response. *arXiv preprint arXiv:2004.11838*, 2020.

Md Zahangir Alom, Tarek M Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Mahmudul Hasan, Brian C Van Essen, Abdul AS Awwal, and Vijayan K Asari. A state-of-the-art survey on deep learning theory and architectures. *Electronics*, 8(3): 292, 2019.

Maneet Singh, Shruti Nagpal, Richa Singh, and Mayank Vatsa. Dual directed capsule network for very low resolution image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 340–349, 2019.

Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Chiman Kwan, Bryan Chou, Jonathan Yang, and Trac Tran. Compressive object tracking and classification using deep learning for infrared videos. In *Pattern Recognition and Tracking XXX*, volume 10995, page 1099506, 2019.

Yao-Hung Hubert Tsai, Nitish Srivastava, Hanlin Goh, and Ruslan Salakhutdinov. Capsules with inverted dot-product attention routing. *arXiv preprint arXiv:2002.04764*, 2020.

Hussein Mouzannar, Yara Rizk, and Mariette Awad. Damage identification in social media posts using multimodal deep learning. In *ISCRAM*. Rochester, NY, USA, 2018b.

Zishan Ahmad, Raghav Jindal, NS Mukuntha, Asif Ekbal, and Pushpak Bhattachharyya. Multi-

modality helps in crisis management: An attention-based deep learning approach of leveraging text for image classification. *Expert Systems with Applications*, 195:116626, 2022.

Larry J King. Social media use during natural disasters: An analysis of social media usage during hurricanes harvey and irma. 2018.

Yusuf Selman Inanc. Turkey earthquake: Trapped victims cry for help on social media. https://www.middleeasteye.net/news/turkey-earthquake-trapped-victims\protect\discretionary{\char\hyphenchar\font}{}{}cry-help-social-media, 2023. Accessed on 14 February 2023.

Sourasekhar Banerjee, Yashwant Singh Patel, Pushkar Kumar, and Monowar Bhuyan. Towards post-disaster damage assessment using deep transfer learning and gan-based data augmentation. In *24th International Conference on Distributed Computing and Networking*, pages 372–377, 2023.

Edy Irwansyah, Hansen Young, and Alexander AS Gunawan. Multi disaster building damage assessment with deep learning using satellite imagery data. *International Journal of Intelligent Systems and Applications in Engineering*, 11(1):122–131, 2023.

Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1422–1431, 2022.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54 (10s):1–41, 2022.

Chen Zhao, Renjun Shuai, Li Ma, Wenjia Liu, and Menglin Wu. Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit. *Multimedia tools and applications*, 81(17):24265–24300, 2022.

Kang An and Yanping Zhang. Lpvit: A transformer based model for pcb image classification and defect detection. *IEEE Access*, 10:42542–42553, 2022.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Huali Xu, Shuaifeng Zhi, Shuzhou Sun, Vishal M Patel, and Li Liu. Deep learning for cross-domain few-shot visual recognition: A survey. *arXiv preprint arXiv:2303.08557*, 2023a.

Yilmaz Korkmaz, Salman UH Dar, Mahmut Yurt, Muzaffer Özbey, and Tolga Cukur. Unsupervised mri reconstruction via zero-shot learned adversarial transformers. *IEEE Transactions on Medical Imaging*, 41(7):1747–1763, 2022.

De Cheng, Gerong Wang, Bo Wang, Qiang Zhang, Jungong Han, and Dingwen Zhang. Hybrid routing transformer for zero-shot learning. *Pattern Recognition*, 137:109270, 2023.

Sayak Nag, Orpaz Goldstein, and Amit K Roy-Chowdhury. Semantics guided contrastive learning of transformers for zero-shot temporal activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6243–6253, 2023.

Chengming Xu, Siqian Yang, Yabiao Wang, Zhanxiong Wang, Yanwei Fu, and Xiangyang Xue. Exploring efficient few-shot adaptation for vision transformers. *arXiv preprint arXiv:2301.02419*, 2023b.

Bo Jiang, Kangkang Zhao, and Jin Tang. Rgtransformer: Region-graph transformer for image representation and few-shot classification. *IEEE Signal Processing Letters*, 29:792–796, 2022.

Yishu Peng, Yaru Liu, Bing Tu, and Yuwen Zhang. Convolutional transformer-based few-shot learning for cross-domain hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, and Ming-Hsuan Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *Proceedings of the 17th European Conference on Computer Vision–ECCV 2022, Part XIII*, pages 37–54, Tel Aviv, Israel, 2022. Springer.

Xiyu Wang, Pengxin Guo, and Yu Zhang. Domain adaptation via bidirectional cross-attention transformer. *arXiv preprint arXiv:2201.05887*, 2022.

Maryam Sultana, Muzammal Naseer, Muhammad Haris Khan, Salman Khan, and Fahad Shahbaz Khan. Self-distilled vision transformer for domain generalization. In *Proceedings of the Asian Conference on Computer Vision*, pages 3068–3085, 2022.

Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7191–7200, 2022.

Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 520–530, 2023.

Soudabeh Taghian and Doina Caragea. Disaster image classification using capsule networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

Zhihao Ma, Mengke Yuan, Jiaming Gu, Weiliang Meng, Shibiao Xu, and Xiaopeng Zhang. Triple-

strip attention mechanism-based natural disaster images classification and segmentation. *The Visual Computer*, 38(9-10):3163–3173, 2022.

Rani Koshy and Sivasankar Elango. Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. *Neural Computing and Applications*, 35 (2):1607–1627, 2023.

Zhihao Ma, Wei Li, Muyang Zhang, Weiliang Meng, Shibiao Xu, and Xiaopeng Zhang. Htcvit: an effective network for image classification and segmentation based on natural disaster datasets. *The Visual Computer*, pages 1–13, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021.

Iustin Sirbu, Tiberiu Sosea, Cornelia Caragea, Doina Caragea, and Traian Rebedea. Multimodal semi-supervised learning for disaster tweet classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2711–2723, 2022.

Saideshwar Kotha, Smitha Haridasan, Ajita Rattani, Aaron Bowen, Glyn Rimmington, and Atri

Dutta. Multimodal combination of text and image tweets for disaster response assessment. International Workshop on Data-driven Resilience Research, 2022.

Akash Kumar Gautam, Luv Misra, Ajit Kumar, Kush Misra, Shashwat Aggarwal, and Rajiv Ratn Shah. Multimodal analysis of disaster tweets. In *2019 IEEE Fifth international conference on multimedia big data (BigMM)*, pages 94–103. IEEE, 2019.

Li Xukun and Doina Caragea. Improving disaster-related tweet classification with a multimodal approach. In *ISCRAM 2020 Conference Proceedings–17th International Conference on Information Systems for Crisis Response and Management*, 2020.

Sreenivasulu Madichetty, Sridevi Muthukumarasamy, and P Jayadev. Multi-modal classification of twitter data during disasters for humanitarian response. *Journal of ambient intelligence and humanized computing*, 12:10223–10237, 2021.

Qi Chen, Wei Wang, Kaizhu Huang, Suparna De, and Frans Coenen. Multi-modal adversarial training for crisis-related data classification on social media. In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 232–237. IEEE, 2020.

Camila E Young, Camila E Young, Erica D Kuligowski, and Aashna Pradhan. *A review of social media use during disaster response and recovery phases*. US Department of Commerce, National Institute of Standards and Technology, 2020.

Ashir Ahmed. Use of social media in disaster management. 2011.

Carlos Villegas, Matthew Martinez, and Matthew Krause. Lessons from harvey: Crisis informatics for urban resilience. *Rice University Kinder Institute for Urban Research*, 2018.

Marina Koren. Using twitter to save a newborn from a flood. The Atlantic, 2017. URL https://www.theatlantic.com/technology/archive/2017/08/harvey-rescue-twitter/538191/.

Firoj Alam, Ferda Ofli, Muhammad Imran, and Michael Aupetit. A twitter tale of three hurricanes: Harvey, Irma, and Maria. *arXiv preprint arXiv:1805.05144*, 2018c.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: Survey summary. In *Companion Proceedings of the The Web Conference 2018*, pages 507–511, 2018.

Reza Mazloom, Hongmin Li, Doina Caragea, Cornelia Caragea, and Muhammad Imran. A hybrid domain adaptation approach for identifying crisis-relevant tweets. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 11(2):1–19, 2019.

Soudabeh Taghian and Doina Caragea. Disaster image classification using pre-trained transformer and contrastive learning models. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–11. IEEE, 2023b.

Yujia Wu, Jing Li, Jia Wu, and Jun Chang. Siamese capsule networks with global and local features for text classification. *Neurocomputing*, 390:88–98, 2020.

Bowen Zhang, Xiaofei Xu, Min Yang, Xiaojun Chen, and Yunming Ye. Cross-domain sentiment classification by capsule network with semantic rules. *IEEE Access*, 6:58284–58294, 2018.

Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538*, 2018.

Zhe Chen, Shang Li, Lin Ye, and Hongli Zhang. Multi-label classification of legal text based on label embedding and capsule network. *Applied Intelligence*, 53(6):6873–6886, 2023.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In *Proceedings of*

the 15th Annual Conference for Information Systems for Crisis Response and Management (ISCRAM), 2018.

Ashis Kumar Chanda. Efficacy of bert embeddings on predicting disaster from twitter data. *arXiv preprint arXiv:2108.10698*, 2021.

AK Ningsih and AI Hadiana. Disaster tweets classification in disaster response using bidirectional encoder representations from transformer (bert). In *IOP Conference Series: Materials Science and Engineering*, volume 1115, page 012032. IOP Publishing, 2021.

Hamada M Zahera, Ibrahim A Elgendy, Rricha Jalota, and Mohamed Ahmed Sherif. Fine-tuned bert model for multi-label tweets classification. In *TREC*, pages 1–7, 2019.

Warih Maharani. Sentiment analysis during jakarta flood for emergency responses and situational awareness in disaster management using bert. In *2020 8th International Conference on Information and Communication Technology (ICoICT)*, pages 1–5. IEEE, 2020.

Sumera Naaz, Zain Ul Abedin, and Danish Raza Rizvi. Sequence classification of tweets with transfer learning via bert in the field of disaster management. *EAI Endorsed Transactions on Scalable Information Systems*, 8(31):e8, 2021.

Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. Cross-lingual disaster-related multi-label tweet classification with manifold mixup. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 2020.

Guoqin Ma. Tweets classification with bert in the field of disaster management. *StudentReport@ Stanford. edu*, 2019.

Mensah Kwabena Patrick, Adebayo Felix Adekoya, Ayidzoe Abra Mighty, and Baagyire Y Edward. Capsule networks–a survey. *Journal of King Saud University - computer and information sciences*, 34(1):1295–1310, 2022.

P Demotte, K Wijegunarathna, D Meedeniya, and I Perera. Enhanced sentiment extraction architecture for social media content analysis using capsule networks. *Multimedia Tools and Applications*, pages 1–26, 2021.

Akma Khikmatullaev. *Capsule Neural Networks for Text Classification*. Universit¨at Bonn, 2019.

Min Yang, Wei Zhao, Lei Chen, Qiang Qu, Zhou Zhao, and Ying Shen. Investigating the transferring capability of capsule networks for text classification. *Neural Networks*, 118:247–261, 2019.

Jaeyoung Kim, Sion Jang, Eunjeong Park, and Sungchul Choi. Text classification using capsules. *Neurocomputing*, 376:214–221, 2020.

Yi Zhao, Yanyan Shen, and Junjie Yao. Recurrent neural network for text classification with hierarchical multiscale dense connections. In *IJCAI*, pages 5450–5456, 2019b.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Gang Liu and Jiabao Guo. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338, 2019.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

Guangquan Lu, Jiangzhang Gan, Jian Yin, Zhiping Luo, Bo Li, and Xishun Zhao. Multi-task learning using a hybrid representation for text classification. *Neural Computing and Applications*, 32(11):6467–6480, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829, 2021.

Rohit Kumar Kaliyar. A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of bert. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 336–340. IEEE, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Hongmin Li, Nicolais Guevara, Nic Herndon, Doina Caragea, Kishore Neppalli, Cornelia Caragea, Anna Cinzia Squicciarini, and Andrea H Tapia. Twitter mining for disaster response: A domain adaptation approach. In *ISCRAM*, 2015.

Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. Cross-language domain adaptation for classifying crisis-related short messages. *arXiv preprint arXiv:1602.05388*, 2016b.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. Empirical study of zero-shot ner with chatgpt. *arXiv preprint arXiv:2310.10035*, 2023.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2): 1–40, 2023.

Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. Information retrieval meets large language models: a strategic report from chinese ir community. *AI Open*, 4:80–90, 2023.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493, 2024.

Salman Rahman, Lavender Yao Jiang, Saadia Gabriel, Yindalon Aphinyanaphongs, Eric Karl Oermann, and Rumi Chunara. Generalization in healthcare ai: Evaluation of a clinical large language model. *arXiv preprint arXiv:2402.10965*, 2024.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer, 2024.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. Collaborative large language model for recommender systems. In *Proceedings of the ACM on Web Conference 2024*, pages 3162–3172, 2024.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*, 2024.

Rabbia Ahmed, Sadaf Abdul Rauf, and Seemab Latif. Leveraging large language models and prompt settings for context-aware financial sentiment analysis. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–9. IEEE, 2024.

Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. Benchmarking large language models for molecule prediction tasks. *arXiv preprint arXiv:2403.05075*, 2024.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Yingjie Hu, Gengchen Mai, Chris Cundy, Kristy Choi, Ni Lao, Wei Liu, Gaurish Lakhanpal, Ryan Zhenqi Zhou, and Kenneth Joseph. Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 37(11):2289–2318, 2023.

Vinicius G Goecks and Nicholas R Waytowich. Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios. *arXiv preprint arXiv:2306.17271*, 2023.

Grace Colverd, Paul Darm, Leonard Silverberg, and Noah Kasmanoff. Floodbrain: Flood disaster reporting by web-based retrieval augmented generation with an llm. *arXiv preprint arXiv:2311.02597*, 2023.

Abdul Wahab Ziaullah, Ferda Ofli, and Muhammad Imran. Monitoring critical infrastructure facilities during disasters using large language models. *arXiv preprint arXiv:2404.14432*, 2024.