

Comparative Study of Deep Learning Models for Unimodal & Multimodal Disaster Data for Effective Disaster Management

دراسة مقارنة لنماذج التعلم العميق لبيانات الكوارث الأحادية الواسطة والمتعددة
الوسائط من أجل الإدارة الفعالة للكوارث

by

DENA AHMED MOHAMED

**Dissertation submitted in fulfilment
of the requirements for the degree of
MSc INFORMATICS**

at

The British University in Dubai

July 2021

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.



Signature of the student

COPYRIGHT AND INFORMATION TO USERS

The author whose copyright is declared on the title page of the work has granted to the British University in Dubai the right to lend his/her research work to users of its library and to make partial or single copies for educational and research use.

The author has also granted permission to the University to keep or make a digital copy for similar use and for the purpose of preservation of the work digitally.

Multiple copying of this work for scholarly purposes may be granted by either the author, the Registrar or the Dean only.

Copying for financial gain shall only be allowed with the author's express permission.

Any use of this work in whole or in part shall respect the moral rights of the author to be acknowledged and to reflect in good faith and without detriment the meaning of the content, and the original authorship.

Abstract

Multimodal data of text and images on social media posts hold valuable information that can be utilized during crisis events. Such information includes requests for help, rescue efforts, warnings, infrastructure damage, missing people, injured or dead individuals, volunteers, donations, and many more. Many studies focus only on the text modalities, single classification tasks and small-scale home-grown datasets when studying how useful social media data can be for emergency services. In this study, a multimodal deep learning system for automatic classification of disaster tweets was developed. Two classification tasks were tackled, which are informativeness and the humanitarian category. An extensive comparison between unimodal text-only, unimodal image-only and multimodal deep learning models across three different representative disaster datasets (CrisisMMD, CrisisNLP, and CrisisLex26) was done. Convolutional neural networks are utilized for defining the deep learning architectures. Experiments across the multiple settings and datasets show that multimodal models perform better than their unimodal counterparts. It was also found that mapping between the diverse humanitarian categories and consolidating smaller datasets with larger ones significantly improves the models' performance when compared to individual datasets. The consolidated dataset can serve as a new baseline multimodal dataset for further research directions.

خلاصة

تحتوي البيانات المتعددة الوسائط للنصوص والصور على منشورات وسائل التواصل الاجتماعي على معلومات قيمة يمكن استخدامها أثناء الأزمات والكوارث. وتشمل هذه المعلومات طلبات المساعدة، جهود الإنقاذ، التحذيرات، أضرار البنية التحتية، المفقودين، الجرحى أو القتلى، المتطوعين، التبرعات، وغيرها الكثير. تركز العديد من الدراسات فقط على طرائق النص، ومهام التصنيف الفردية، ومجموعات البيانات المحلية الصغيرة عند دراسة مدى فائدة بيانات وسائل التواصل الاجتماعي لخدمات الطوارئ. في هذه الدراسة، تم تطوير نظام تعلم عميق متعدد الوسائط للتصنيف التلقائي لتغريدات الكوارث. تم تناول مهمتين للتصنيف، هما المعلوماتية والفئة الإنسانية. تم إجراء مقارنة شاملة بين نماذج التعلم العميقة للنصوص الأحادية الوسائط فقط، والصورة الأحادية الوسائط فقط، والنماذج المتعددة الوسائط عبر ثلاث مجموعات بيانات تمثيلية مختلفة عن الكوارث (CrisisMMD, CrisisNLP, CrisisLex26). تستخدم الشبكات العصبية الملتفة في تحديد هياكل التعلم العميق. وتبين التجارب عبر الإعدادات المتعددة ومجموعات البيانات أن أداء نماذج النقل المتعدد الوسائط أفضل من أداء نظيراتها الأحادية الوسائط. ووجد أيضا أن توحيد مختلف الفئات الإنسانية ودمج مجموعات البيانات الأصغر حجما مع مجموعات البيانات الأكبر حجما يحسن بشكل كبير أداء النماذج إذا ما قورنت بمجموعات البيانات الفردية. ويمكن أن تكون مجموعة البيانات الموحدة بمثابة مجموعة بيانات أساسية جديدة متعددة الوسائط لمزيد من التوجيهات البحثية.

ACKNOWLEDGMENT

I would like to express my appreciation and gratitude to my supervisor Professor Sherief Abdullah for his valuable guidance and support throughout the dissertation. I would also like to thank Professor Khaled Shaalan for the knowledge taught during the Masters journey.

Finally, my sincere thanks to my family for their continuous support.

Table of Contents

| | |
|--|-----------|
| Chapter 1 : Introduction..... | 1 |
| 1.1. Motivation and Problem Statement | 2 |
| 1.2. Aims and Objectives | 4 |
| 1.3. Research Questions..... | 5 |
| 1.4. Research Methodology | 5 |
| 1.5. Dissertation Structure | 7 |
| Chapter 2: Background and Related Work..... | 8 |
| 2.1. Disasters | 9 |
| 2.1.1. Types of Disasters | 9 |
| 2.1.2. Disaster Management Cycle..... | 12 |
| 2.1.3. Disaster Management Success Factors..... | 15 |
| 2.2. Social Media and Crisis Situations..... | 16 |
| 2.2.1. Social Media Usage during Disasters..... | 19 |
| 2.2.2. Twitter during disasters | 24 |
| 2.3. Machine Learning for Emergency Response | 28 |
| 2.3.1. Deep Learning..... | 29 |
| 2.3.2. Disaster Social Media Text Classification | 31 |
| 2.3.3. Disaster Social Media Image Classification..... | 34 |
| 2.3.4. Disaster Social Media: Multimodal Data | 36 |
| 2.4. Summary | 39 |
| Chapter 3: Methodology | 41 |
| 3.1. Overall System Architecture | 41 |
| 3.2. Datasets | 44 |
| 3.2.1. First Dataset: CrisisMMD | 44 |
| 3.2.2. Second Dataset: CrisisNLP | 47 |
| 3.2.3. Third Dataset: CrisisLex26 | 60 |
| 3.3. Data Preprocessing | 66 |
| 3.4. Classification Approach | 67 |
| 3.4.1. Text Modality | 67 |

| | |
|--|------------|
| 3.4.2. Image Modality | 69 |
| 3.4.3. Multimodal Classification | 72 |
| Chapter 4: Results and Discussion..... | 73 |
| 4.1. Performance Evaluation Metrics | 73 |
| 4.2. Models' Performance for CrisisMMD | 76 |
| 4.3. Models' Performance for CrisisNLP | 78 |
| 4.4. Models' Performance for CrisisLex26 | 81 |
| Chapter 5: Conclusions and Future Work..... | 86 |
| 5.1. Conclusion..... | 86 |
| 5.2. Limitations and Future Work | 89 |
| References | 92 |
| Appendix A | 107 |

List of Figures

| | |
|---|----|
| Figure 1: Impacts of Disasters 1980-1999 vs. 2000-2019 | 2 |
| Figure 2: Research Areas for Surveyed Literature | 8 |
| Figure 3: Number of Disasters Per Year from 2010 till 2021 | 11 |
| Figure 4: Total Disaster Events by Type: 1980-1999 vs. 2000-2019 | 12 |
| Figure 5: Disaster Management Cycle | 13 |
| Figure 6: Global Social Media Usage Statistics | 18 |
| Figure 7: Overall Architecture of Multimodal Classification Approach..... | 42 |
| Figure 8: VGG-16 Model Architecture | 69 |

List of Tables

| | |
|---|-----|
| Table 1: Disaster Types | 10 |
| Table 2: Sample Labelled Disaster Tweets | 44 |
| Table 3: CrisisMMD Informativeness Distribution | 45 |
| Table 4: CrisisMMD Humanitarian Category Distribution..... | 45 |
| Table 5: Informativeness Data Split over CrisisMMD Subset | 47 |
| Table 6: Humanitarian Category Data Split over CrisisMMD Subset | 47 |
| Table 7: CrisisNLP Disaster Collections Distribution | 48 |
| Table 8: CrisisNLP Disaster Distribution for Labelled Tweets | 56 |
| Table 9: CrisisNLP Disaster Distribution After Filtering | 58 |
| Table 10: CrisisNLP Informativeness Distribution | 58 |
| Table 11: CrisisNLP Humanitarian Category Distribution | 59 |
| Table 12: Informativeness Data Split over CrisisNLP Dataset | 59 |
| Table 13: Humanitarian Category Data Split over CrisisNLP Dataset | 60 |
| Table 14: CrisisLex26 Disaster Distribution for Labelled Tweets..... | 64 |
| Table 15: : CrisisLex26 Informativeness Distribution | 65 |
| Table 16: CrisisLex26 Humanitarian Category Distribution | 65 |
| Table 17: Informativeness Data Split over CrisisLex26 Dataset | 66 |
| Table 18: Humanitarian Category Data Split over CrisisLex26 Dataset..... | 66 |
| Table 19: Machine Learning Python Libraries used | 76 |
| Table 20: Results for informativeness classification task for CrisisMMD | 76 |
| Table 21: Results for humanitarian category classification task for CrisisMMD | 77 |
| Table 22: Results for informativeness classification task for CrisisNLP | 78 |
| Table 23: Results for humanitarian category classification task for CrisisNLP | 79 |
| Table 24: Results for informativeness classification task for CrisisNLP+CrisiMMD | 80 |
| Table 25: Results for humanitarian category task for CrisisNLP+CrisisMMD | 81 |
| Table 26: Results for informativeness classification task for CrisisLex26 | 82 |
| Table 27: Results for humanitarian category classification task for CrisisLex26 | 82 |
| Table 28: Results for informativeness classification task for Consolidated Dataset (CrisisMMD+CrisisNLP+CrisisLex26) | 83 |
| Table 29: Results for humanitarian category task for Consolidated Dataset (CrisisMMD+CrisisNLP+CrisisLex26) | 84 |
| Table 30: Machine Learning Python Libraries in both studies | 108 |

Chapter 1

Introduction

This chapter presents an overview of the representation of disasters and crisis events in social media and how this data can be utilized for effective disaster management. Moreover, it presents the motivations for the dissertation, research questions, research methodology overview, and contributions to the current body of knowledge.

1.1. Motivation and Problem Statement

During the past twenty years, disasters resulted in around 1.23 million deaths (with an average of 60 thousand deaths annually), and affected approximately four billion individuals. Moreover, they resulted in economic losses of almost three trillion USD all over the world. These damages are much greater than those in the earlier twenty years from 1980 to 1999, which shows that the damages caused by disasters are increasing over the years (CRED 2020). These figures are shown in Figure 1.

Disasters and crisis events, whether natural or man-made, result in significant damage to societies. Whether it is an earthquake, flood, fire, tsunami, bombing, pandemic, or any other disaster type, societies must be prepared to face such situations by employing effective disaster management strategies to reduce the risks, mitigate the damage inflicted and recover from the crisis's aftermath (Keim & Noji 2011). Proper communication of information is an essential factor for efficient emergency response operations. When a disaster strikes, people

are frantically looking for information to grasp the gravity of the situation and understand what is going on to behave accordingly to avoid risk and stay safe (Williams, Williams & Burton 2012). One of the primary sources for such timely disaster-relevant information is social media platforms (Simon, Goldberg & Adini 2015).

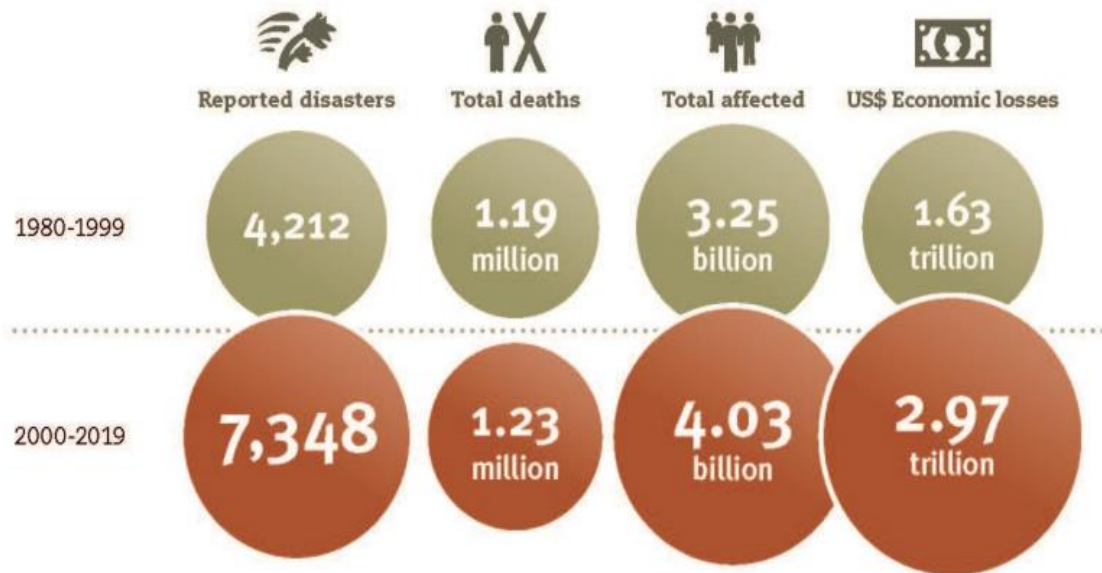


Figure 1: Impacts of Disasters 1980-1999 vs. 2000-2019 (CRED 2020)

Social media is an integral part of our everyday lives, with people checking them all day long, especially with the convenient portability of mobile devices. This has led to having billions of people around the world sharing, posting, interacting, and consuming content online (Kemp 2021). With such a tremendous amount of information available on social media platforms, a valuable information source becomes available for emergency services. Disaster-related information can be gathered from social media and analyzed to extract useful information that would assist with emergency response efforts. The use of social media as a medium for populating disaster-related information to facilitate emergency response has been gaining momentum in recent years (especially since 2010). There has

been an exponential increase in people's participation during crisis events including posting, sharing, gathering, and spreading information to assist the emergency response operations (Mills et al. 2009; Fraustino, Brooke & Yan 2012; Imran et al. 2013). So, when a disaster occurs, individuals used social media intensively to post and share information (text, images or videos) relevant to the crisis event such as missing people, injured or dead individuals, affected people, damages to infrastructure, rescue requests, pleas for help, warnings, donation calls, volunteering requests, emotional support and much more. Several studies show that such information is beneficial for emergency response services and concerned humanitarian organizations to improve the disaster management motion through increased situational awareness and more informed planning for rescue operations (Ilyas 2014; Landwehr et al. 2016; Panagiotopoulos et al. 2016; Laylavi, Rajabifard & Kalantari 2017).

With such a large volume of unstructured content in disaster social media data, challenges arise during analysis. These challenges include information overload, noisy data, and filtering irrelevant data. Several AI techniques for processing social media data have been presented in recent years to address these challenges. Studies have experimented with classic machine learning algorithms and deep-learning based approaches as well when analyzing such data to extract useful information. The most popular classification tasks with regards to disaster social media data are informativeness and humanitarian category. Informativeness detects if the disaster tweet is informative or not to determine if it can provide relevant crisis-related information that would help emergency workers or not. The humanitarian category task classifies disaster social media data into categories such as

injured people, help requests, affected individuals, donations, warnings, damage, etc. Such classification helps narrow down the immense amount of disaster data available so that emergency services can better manage their resources for timely response to those in need corresponding to the severity of the case. So, for instance, when a disaster first hits, content in help requests and injured people categories would have the highest priority and can assist with efficient rescue operations (Landwehr et al. 2016; Jomaa, Rizk & Awad 2017; Laylavi, Rajabifard & Kalantari 2017).

1.2. Aims and Objectives

This study aims to build a multimodal deep learning system for automatic classification of disaster tweets. The classification tasks will be informativeness (informative vs. not informative) and the humanitarian category. It is also a comparative study for unimodal and multimodal deep learning models across multiple representative disaster datasets. The study also tackles the issue of inconsistency of classes and structure when using multiple datasets by mapping. It will also create consolidated datasets to compensate for the lower performance of small datasets in deep learning models. The system performance between the multiple disaster data sets with different settings is to be compared. This will provide insight into which system setting generates the best performance when classifying disaster tweets and can be the most beneficial to emergency service workers for efficient disaster management.

1.3. Research Questions

The goal of this dissertation is to utilize social media disaster posts to support more efficient disaster management by developing a multimodal deep learning system to automatically classify disaster data over informativeness and humanitarian category tasks and do a comparative analysis of unimodal/multimodal deep learning models over different representative disaster datasets.

The following are the research questions:

1. Is it possible to integrate multiple disaster datasets even if the labels are not identical in all of them? If possible, how effective will the integration be?
2. How does performance of unimodal and multimodal models compare across different disaster datasets?

1.4. Research Methodology

Since social media data is multimodal in nature, containing a mixture of text and images, multimodal models were found to produce better results when compared to unimodal models dealing with either text-only or image-only data. This dissertation will develop deep learning models for the classification of unimodal and multimodal disaster data. A Convolutional Neural Network is used for unimodal models with text-only data, and a VGG16 network is used for unimodal models with image-only data. The multimodal model implements feature fusion, where it obtains two feature vectors from both the text modality and the image modality. For both classification tasks informativeness and humanitarian category, three models will be built: unimodal with text-only data, unimodal with image-only data, and

multimodal with both text and image modalities. This process will be done for three representative disaster datasets which are CrisisMMD, CrisisNLP, and CrisisLex26. The classification performance will be compared across all three datasets.

Mapping will also be done between the different humanitarian categories of the datasets for uniform classification results. One challenging point with available labelled disaster datasets is that the humanitarian categories are so diverse, with many different variations and namings. So, for instance, one dataset could have one category for injured/missing/dead people and another dataset could have those as three separate categories. Therefore, mapping between the different categories has been done across the three datasets to have uniform classes across all of them.

In addition, the CrisisMMD dataset, being the leading multimodal dataset in the Crisis Informatics scene, has the largest amount of multimodal data compared to CrisisNLP and CrisisLex26. After the mapping, data from the larger CrisisMMD was consolidated with the other two datasets, leading to significant improvement in classification results compared to when only the smaller datasets were used in training the models. The three models (two unimodal and one multimodal) are evaluated across the three datasets to ensure the accuracy and establish their reliability. The work done also emphasizes and further supports previous findings that multimodal models perform better than unimodal models.

Furthermore, most of the studies to be surveyed in the literature review in the next chapter show research done on small-scale datasets, most of which are home-grown for only specific disaster types. This results in models that will not generalize well when faced with different kinds of disasters. This presents the need for having a large dataset that would include a

variety of disaster types with multimodal information. Therefore, this dissertation merges the three representative baseline disaster datasets into a consolidated multimodal dataset with a very diverse representation of disasters of many types including earthquakes, typhoons, floods, fires, etc.

1.5. Dissertation Structure

This dissertation is organized as follows. Chapter 2 presents a comprehensive literature review about disasters and crisis events with their various types, disaster management cycle, social media and its importance in our everyday lives in general and with respect to disaster response specifically, how disaster social media is presented, and machine learning approaches for utilizing such data for effective disaster response. Next, chapter 3 demonstrates the methodology of this study, datasets used, detailed phases of the mapping and comparative analysis performed between the three representative datasets, and the classification tasks performed using deep learning models. Chapter 4 then presents the results and findings from the classification tasks for all datasets, and compares the performance metrics for several cases and situations for unimodal and multimodal models. Finally, chapter 5 concludes by summarizing this work's findings and sheds some light on possible future work in this research direction.

Chapter 2

Background and Related Work

The content of this dissertation falls under the general umbrella of Crisis Informatics, which can be defined broadly as the integration of informational, technical, and social aspects of crisis events and how they are all interconnected (Reuter & Kaufhold 2018) (Tan et al. 2017). The term was formulated by (Hagar 2010), and then further developed by (Palen et al. 2009). Crisis Informatics “is a multidisciplinary field that combines social science knowledge of disasters together with computing; where its central principle is that people use communication technology and personal information to respond to disasters/crisis events in creative ways to deal with uncertainty” (Reuter, Hughes & Kaufhold 2018) (Palen & Anderson 2016).

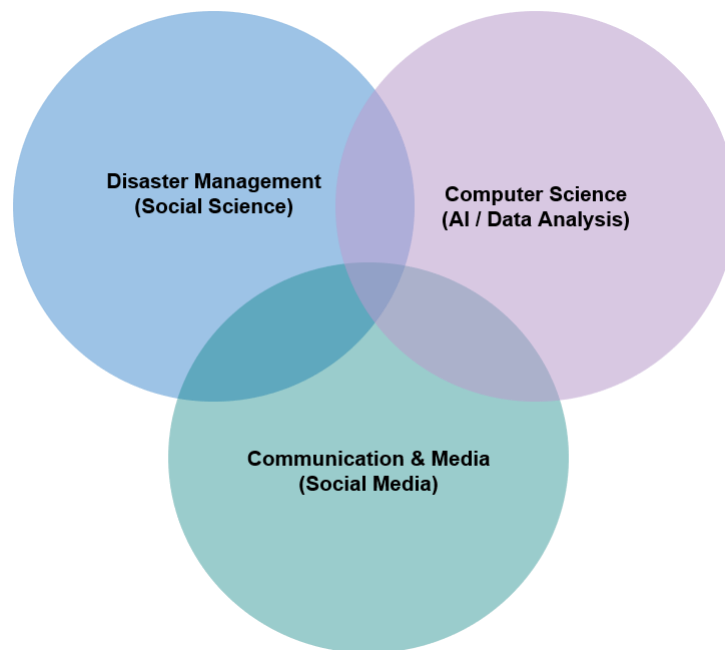


Figure 2: Research Areas for Surveyed Literature

This literature review will demonstrate background information and related studies from the multiple research domains of social media, disaster management, and AI for data analysis.

Figure 2 shows these involved research areas.

2.1. Disasters

A disaster refers to either a natural or man-made event that causes severe damage and loss, resulting in temporary paralysis of the response capability of people, communities, organizations, and nations (Ray & Bala 2020). Disasters of different types have a negative impact on environments and need to be managed properly by emergency services. Such negative impacts and disaster management cycles will be discussed in more detail in the following sections.

2.1.1. Types of Disasters

Natural disasters are crisis events that resulted from natural causes such as floods, hurricanes, earthquakes, etc. These are disasters that humans have no control over. On the other hand, man-made disasters are the calamitous crisis events that are the direct cause of human behavior and actions such as vehicle accidents, building collapses, wars, etc. (Shaluf 2007). Table 1 shows the subcategories within each disaster type.

Disasters of all types result in severe damages and losses. To demonstrate the degree of such catastrophic effects, some global statistics were gathered.

| Disaster Type | Disaster Subcategory | Disaster Name |
|--------------------|---------------------------|--|
| Natural Disasters | Below Earth's Surface | Tsunami |
| | | Earthquake |
| | | Volcano Eruption |
| | On Earth's Surface | Avalanche |
| | | Landslide |
| | Hydrological/Metrological | Typhoon |
| | | Cyclone |
| | | Hurricane |
| | | Tornado |
| | | Snowstorm |
| | | Flood |
| | | Drought |
| | | Heat/cold waves |
| | Biological | Epidemic |
| | | Pandemic |
| | | Infestations (pest swarms...) |
| Man-made Disasters | Accidents | Plane crash |
| | | Train crash |
| | | Shipwreck |
| | | Car crash |
| | | Building collapse |
| | | Explosion |
| | | Fire |
| | | Chemical (poisoning, pollution, oil spill) |
| | Warfare | Nuclear |
| | | Chemical |
| | | Biological |
| | | Siege |
| | | Blockade |
| | | War between armies |
| | | Civil war |
| | | Terrorist attack |
| | | Bomb Threat |

Table 1: Disaster Types

From 2000 to 2019, over 7000 disasters of various types globally were recorded by EM-DAT (Emergency Events Database). EM-DAT is a global database recording detailed information about natural and man-made disasters including effects, losses, locations, types, and many more disasters from 1900 till now. This database is maintained and owned by the Center of Research on the Epidemiology of Disasters (CRED) in Belgium and was supported by the World Health Organization (WHO) and the Belgian government in inception phases. EM-DAT is considered a leading international database for disaster events, so it is a reliable source (CRED n.d.).

Figure 3 shows a bar chart demonstrating the number of disasters per year from 2010 until 2021. There are almost over 500 disasters annually during the last ten years.

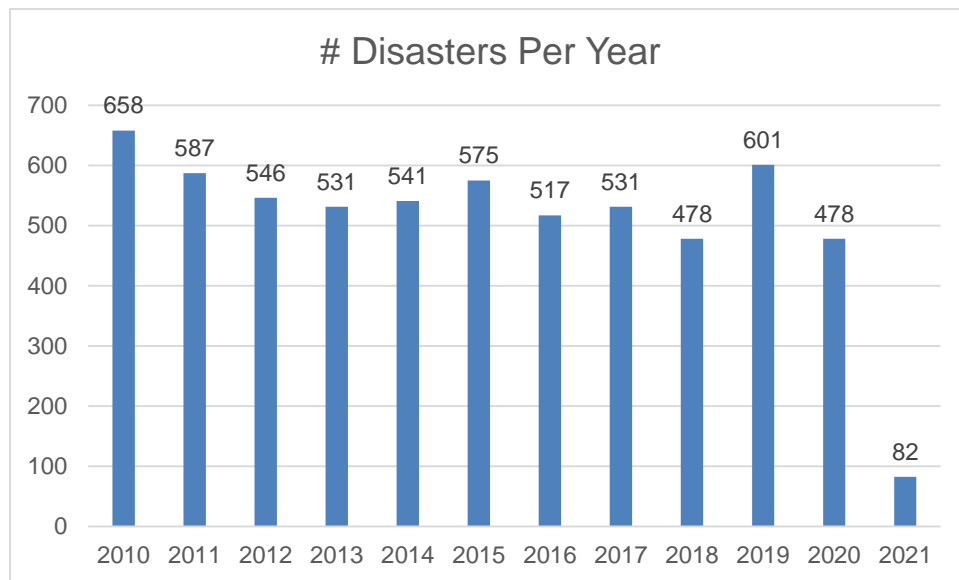


Figure 3: Number of Disasters Per Year from 2010 till 2021 (CRED 2020)

The statistics also showed that between 2000 and 2019, the largest percentage of deaths and affected individuals were because of natural disasters such as floods, storms (including hurricanes, cyclones, tornadoes), earthquakes, extreme temperatures (heat and cold waves), droughts, and wildfires (CRED 2020). This is represented in Figure 4.

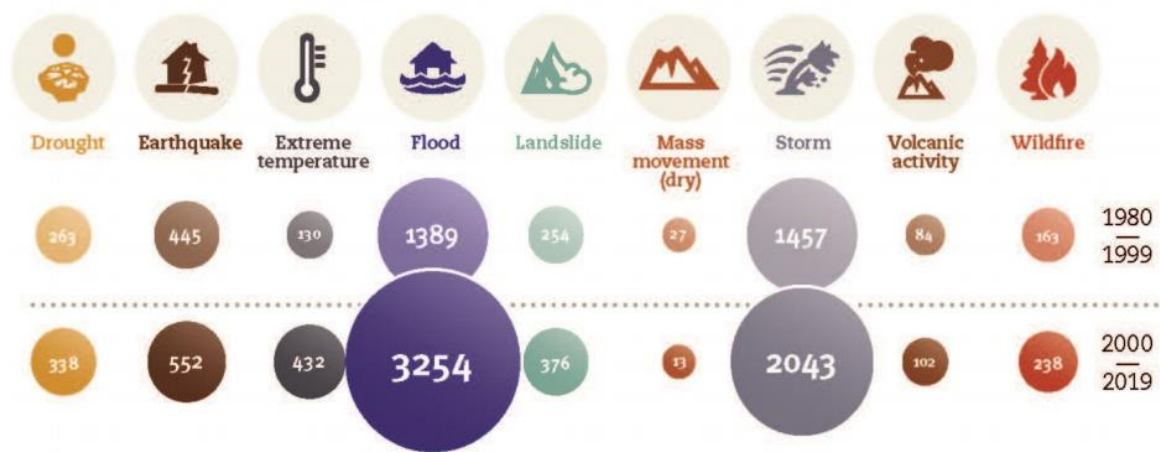


Figure 4: Total Disaster Events by Type: 1980-1999 vs. 2000-2019 (CRED 2020)

All these statistics demonstrate the greatly severe damages and losses caused by disasters, which proves that there is a huge need to strengthen disaster management to reduce possible risks and negative impacts. Comprehensive disaster management systems and emergency response entities are an essential requirement worldwide that should be developed in coordination with related organizations and governments following support guidelines to be able to navigate disaster situations with reduced catastrophic outcomes.

2.1.2. Disaster Management Cycle

Different emergency response organizations tasked with taking action when faced with disasters follow multiple phases to manage crisis events. There are many variations of this

cycle. The most common cycle includes four stages: Mitigation, Preparation, Response, and Recovery) (Nazer et al. 2017). Others follow a more extended version including up to seven stages with the addition of prevention and education. However, the four phases model is the one that is followed by the majority of disaster management organizations all over the world. Sometimes, different names for the stages are made in different countries such as “4 Rs” in New Zealand (Reduction, readiness, response and recovery) (Houston et al. 2015). However, the stages more or less share the same actions. Figure 5 shows the cycle phases.



Figure 5: Disaster Management Cycle

Mitigation is concerned with reducing the chances of disastrous events happening. This stage is usually before the disaster happens, where the concerned emergency response organizations are taking precautions to reduce possible negative impacts of disasters (Houston et al. 2015). An example could be checking the quality of a building’s construction and enhance any areas that could lead to potential future damage. The second phase is

“Preparation”, where the focus is on providing sufficient education, awareness and training for crisis situations so that individuals would be prepared to respond when a disaster occurs (Todd & Todd 2011).

“Response” is the third stage happening immediately after a disaster just happens, and is concerned with handling the urgent threats to affected individuals or entities. The emergency response protocol will be activated, including giving out basic needs such as shelter, medical attention to affected individuals covering injuries and deaths, water, food, search operations, rescue missions, recovery of damage inflicted on infrastructure, and so on (Zhou, Huang & Zhang 2011). The time taken by this stage depends on the nature of the disaster, so that it could take as short as few weeks in case of a small earthquake to months or years for heavy floods. A major obstacle in this response stage is the speed and effectiveness of initiating and managing the emergency response protocol based on the nature of the disaster. Many studies (Noreña et al. 2011) (Zade et al. 2018) (Zhou, Huang & Zhang 2011) have stressed how important it is to collect disaster-relevant information, prioritize responsibilities, and act quickly. The quicker and more efficient the disaster management response is, the less damage inflicted on individuals or communities.

Reports and studies on several disasters (Todd & Todd 2011) also highlighted that the most crucial asset in disaster management is quality information that is available immediately after the disaster happens. Having this information plays a significant role in allowing emergency workers and response organizations to formulate a better-customized emergency response plan and protocol for maximum recovery and damage reduction (Zade et al. 2018). Lack of relevant disaster information causes issues such as incorrect prioritization of rescue

efforts, misleading relief actions or donations out to wrong destinations, thus leading to a dissatisfactory disaster response operation leaving more damage (Zhou, Huang & Zhang 2011). This further emphasizes the importance of gathering relevant actionable information for disasters to enhance the disaster management cycle and have better response results. Years ago, before technological advancements, it was extremely hard to obtain accurate disaster information, especially in major natural disasters with severe damage inflicted. However, nowadays with the Internet and social media available, there are many ways and sources to be utilized for collecting information (to be discussed in more detail in upcoming sections).

The fourth and last stage happens after the disaster, entailing the recovery, restoration and reconstruction of damages resulting from the crisis event. This could include burying the dead, rebuilding damaged structures or buildings, providing financial support to affected entities and all in all bring back a sense of normalcy and stability (Todd & Todd 2011).

2.1.3. Disaster Management Success Factors

Even though different disasters require various relief efforts and needs at different paces, the “Response” stage being the most essential phase, has some crucial success factors that have been established by multiple studies. These factors include proper structuring with wise leadership, accurate situation assessment, and organizing resources depending on accurate estimations of need (Starbird et al. 2010). The component that is shared amongst all these success factors is having actionable accurate disaster-relevant information. Several studies stressed the importance of having quality detailed information for disasters, and how

essential it is in improving the emergency response operations as it will significantly play a huge role in reducing the harm and damage to the affected community (Imran et al. 2013). Since the impact of the disaster is greatly determined by the level of damage to the affected community, it helps close the information gap if individuals from this affected community are gathering such information.

There are different types of information available and useful after a disaster, such as basic needs such as food, water, shelter and medical attention. This information can be obtained from social media since people usually report such requests for help (Bodenhamer 2011). Other information includes missing person reports, deaths, damage reports, and rescue requests. All these types of information give an overview of the entire picture of the aftermath of the disaster event, and help emergency response organizations focus their efforts on those in need for a more effective disaster management operation.

Since social media is a great resource to obtain such information, the upcoming section will discuss literature about social media and how it can be used to obtain useful information for emergency response organizations in disaster management.

2.2. Social Media and Crisis Situations

The earlier sections discussed the importance of information needed by disaster response management organizations for better emergency response, and how that information can be likely gathered from social media. In this section, the focus will be on social media relevant to crisis events, which is one of the research domains of crisis informatics filed as previously shown in Figure 2.

In recent years, social networking websites have strongly influenced the way people interact or communicate in society, which is the reason for their increasing interest in social media for finding information relevant to a disaster (Houston et al. 2015). People began to talk more openly about their life events. Through status updates, they do not have to spend much time discussing and debating their interests. This makes it easy to share information every day, report about the surrounding environment or just chat on social media (Kwon & Han 2013).

Social media is a somewhat general term for various web services and platforms that support networking, allowing users to create content publicly, and communicate with other users' profiles and content (Deller 2011). Of course, these sites can be accessed through various devices such as desktops, laptops, tablets, and smartphones. However, the number of people using mobile devices to access social media platforms is increasing because of the high accessibility.

Looking at some recent statistics compiled in January 2021 from Datareportal's most recent Global Overview report (Kemp 2021), a huge global collection of stats about the digital world, a large increase in the number of Internet and social media users globally can be seen. Within the past year alone, there is an increase of 316 million Internet users, resulting in having 4.66 billion global Internet users. There was also an increase of 93 million mobile phone users, amounting to having 5.22 billion global mobile phone users, which is almost 67% of the world's population. As for social media users, there are currently 4.2 billion worldwide users, which is a growth of 490 million users during the last year alone, which is an enormous growth of over 13% from January 2020 till January 2021. This leaves the

number of total social media users around the globe at over 53% of the world's total population.

Figure 6 also demonstrates more statistics (Kemp 2021) showing that in January 2021, more than half the world's population is using social media. The report also showed that there was an average of around 15 new social media users every second during the past year (January 2020 till January 2021), where there were over 1.3 million new social media users daily around the world.

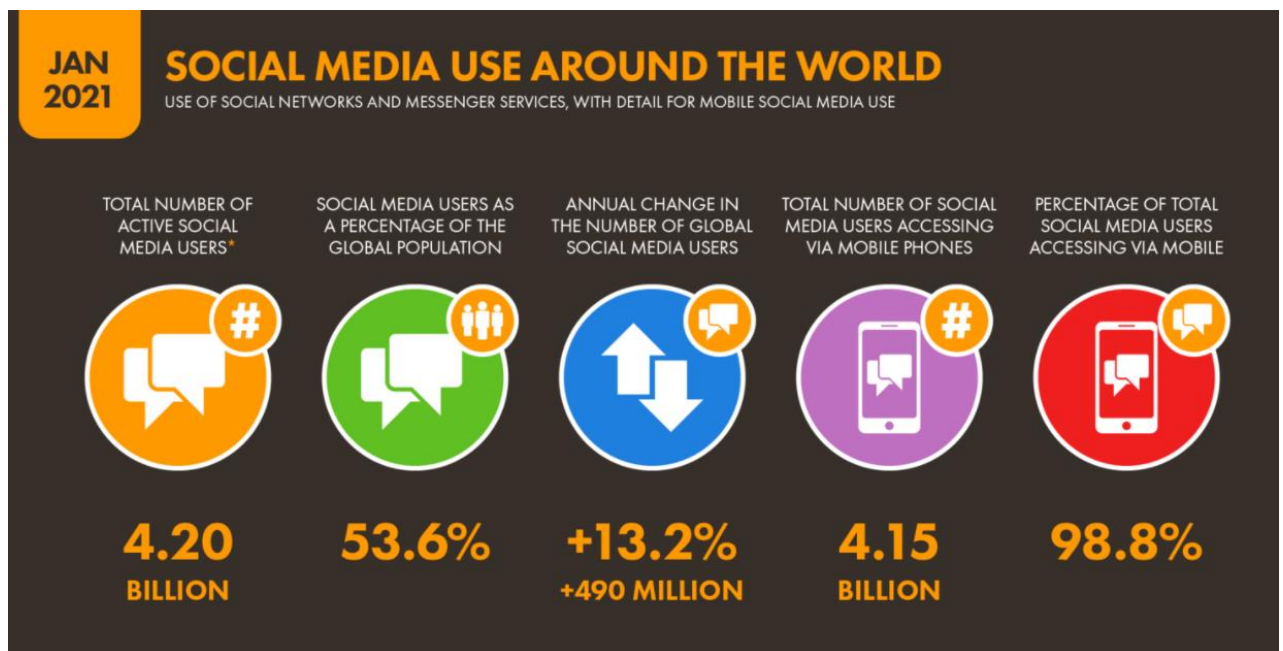


Figure 6: Global Social Media Usage Statistics (Kemp 2021)

As seen from these statistics, social media is an incredibly valuable information source with the effortless accessibility of the Internet, social media and mobile devices in our day and time. The gathering of millions of people has made these social networking sites a source of information for news and research as more and more people join and share information on a daily basis. Social media data is becoming an increasingly essential tool for understanding

human behavior as well. The high value and importance of these social media platforms comes from the fact that millions of users around the world are continuously sharing information regarding all kinds of topics, which makes it a priceless mine of endless real-time data that can be used by all sorts of organizations and research entities (Ray & Bala 2020).

There are multiple categories for social media sites, such as personal networking sites (Facebook, Google+), job-related communities (LinkedIn), discussion forums (Reddit, Quora), blogs (Twitter, WordPress, Blogger, Weibo), image-focused sharing (Instagram, Pinterest, Tumblr), digital content (YouTube, Snapchat, TikTok), social gaming platforms (Discord) and messenger applications (WhatsApp, WeChat, Telegram, Facebook Messenger). Since different social media platforms have various purposes, the information shared on each will be of different forms and thus suitable for various research opportunities (Blank & Reisdorf 2012) (Ray & Bala 2020).

2.2.1. Social Media Usage during Disasters

Many studies (Keim & Noji 2011; Fraustino, Brooke & Yan 2012; Kavanaugh et al. 2012; Williams, Williams & Burton 2012; Martínez-Rojas, Pardo-Ferreira & Rubio-Romero 2018a, 2018b; Zade et al. 2018; Imran et al. 2020) have been made to analyze how people share information on social media during various disasters or crisis events either natural or man-made. All of those studies confirmed that many people depend on social media to get official information during disasters. In addition, people also rely on social media to request for help, share locations or surroundings' updates, coordinate rescue efforts and organize

donations. Social media plays such a critical role in spreading information, increasing awareness, sharing emergency alerts, receiving rescue requests, organizing relief efforts, and many more. The role of using social media for more effective emergency response and disaster management has been seen in multiple cases of past disasters such as the Haiti earthquake in 2010 (Caragea et al. 2011), Japan's earthquake and tsunami in 2011 (Fujii et al. 2011), hurricane sandy in 2012 (Dong, Halem & Zhou 2013), Nepal earthquake in 2015 (Radianti, Hiltz & Labaka 2016), Haiyan typhoon in 2013 (Takahashi, Tandoc & Carmichael 2015), hurricanes Harvey/Irma/Maria in 2017 (Alam et al. 2018), global wildfires (Wang, Ye & Tsou 2016), multiple floods (Basnyat et al. 2017; Li & Di 2017), the spread of Zika virus/Ebola/Covid-19 (Fung et al. 2016; Piedra, Chicaiza & Torres-Guarnizo 2017; Hagen et al. 2018; Mirbabaie et al. 2020; Yu et al. 2020), the Arab Spring (Kumar et al. 2013), Beirut port explosion in 2020 (El Sayed 2020), several gas leaks/oil spills (Muralidharan, Dillistone & Shin 2011; Soares et al. 2020), and many more disasters.

The use of social media during disasters increases immensely since everyone is always looking for the latest updates and needs immediate comprehensive information. For example, a study (Formentin, M., Bortree, D. and Fraustino 2012) examined social media use during the first month of some organizational crisis at a public university, found that students replied almost immediately to the majority of the posts published by the university's official Facebook account page. It also showed that almost around 80 percent of the posts' comments were posted within a 10-15 minutes range, with an increased number of likes and shares than previous posts' engagement in regular times.

A study (Mills et al. 2009) found that the first reports to be made about the earthquake in China back in 2008 were, in fact, from people's tweets on Twitter and not from the government or any other official sources. In another study (Perng et al. 2013), it was found that people posted over 2000 tweets in just the first 30 minutes right before a deadly storm about to hit some festivities in Belgium. That number of tweets exponentially increased to almost eighty thousand tweets throughout the first three to four hours of the crisis alone.

Some studies show that there is a difference between people's usage behavior on social media during disasters and regular times. Moreover, different social media platforms tend to have different engagement patterns for the same disaster. For instance, during the tsunami and earthquake happening in Japan in 2011 (Fujii et al. 2011), there were many worries regarding radiation from destroyed nuclear reactors, and therefore was a very hot topic that was discussed all across the many different social media platforms such as YouTube, Twitter and blogs. Even though it was the same disaster, but the content disseminated on each platform was different. People used YouTube to post, view, and share disaster damage pictures and videos. They used Twitter to obtain the latest disaster-related news and share them, while blogs were used to discuss their feelings and support each other. To manage information shared during disasters and utilize it efficiently, it is essential to know how people use the different social media platforms at that time.

As said before, one of the primary uses of social media is seeking information. With the high level of uncertainty that comes along with any disaster or crisis event, people are more inclined to actively seek new information to stay updated and get real-time updated relevant information. Most of the time, social media is the fastest source of disaster information

(Kavanaugh et al. 2012). It is also often the primary source for such real-time information that is time-sensitive in cases of disasters, because other traditional official media sources are much slower to provide such data (Spiro et al. 2012). For instance, when over five hundred million tweets related to the 2009 Influenza pandemic were analyzed, the data obtained was able to forecast the future rates of influenza spread to high levels of accuracy. These rates were later found to be around 95 percent accurate when compared with the official national statistics that took weeks to formulate from hospitals' reports. In another case in 2007, during the California wildfires disaster (Fraustino, Brooke & Yan 2012), people mostly depended on social media information to know relevant information because the government officials, media agencies and journalists were much slower in providing any useful information (Sutton, Palen & Shklovski 2008).

In a similar sense, social media is used to detect disasters through people's posts. In 2011, during the Virginia earthquake in America, many people learned about the event from reading some tweets about it before feeling it in their geographical location (Houston et al. 2015). A study visualized the movement of the earthquake's waves in comparison to the related disaster tweets, and it was seen that the movement of the tweets was much faster than the earthquake's waves themselves (Honan 2011). This is quite helpful because social media can detect disasters, signal warnings and issue alerts in cases where the pattern of posts somehow indicates the beginnings of a disaster. Early detection of such crisis events will lead to faster emergency response and a more efficient disaster management process in general.

Another important use of social media during disasters is asking for help in urgent situations. In the tsunami and earthquake in Japan in 2011, many tweets containing requests for help (Acar & Muraki 2011). An example tweet “30 people are stuck at Ozaki shrine. The roads are shut down. Anybody, please call the police and fire department”. Other tweets were warnings such as “An alarm of BIG tsunami: Coast of Miyagi prefecture. Escape to any high place”. Moreover, some tweets reported on the individuals’ safety or their surrounding environment such as “The sea level is falling rapidly. I think we’re gonna have tsunami soon”, and “A building exploded. It’s south of Kesennuma-Minami station” (Acar & Muraki 2011). There are also posts on injured, missing and dead people as well. All these information categories are of enormous importance when managing disasters because they are crucial sources to emergency service workers. Through analyzing disaster social media data, emergency workers could be dispatched to help individuals asking for help in certain locations (Keim & Noji 2011). Urgent medical assistance can be provided to injured people, rescue operations can be done for affected individuals, and so on (Taylor, M. et al. 2012). Without such a huge source for such information, emergency services would not know all the locations where help is needed and more damage would be inflicted. This will allow for reaching more people in need and save more lives. People could also know relevant information to take precautions and know updated events as they happen. By sharing such information, public awareness is also raised of the current situation, helping all affected parties.

Disaster social media data can also assist with facilitating donations to affected individuals, and provide interested entities with information regarding people/organizations in need of

such help (Williams et al. 2012). This is also the case for volunteering efforts where sharing such information will help interested volunteers to know which places need their assistance most. For example, after the Haiti earthquake in 2010, it was reported that a large percentage of donations of over 30 million dollars was due to people finding out about the damage and donation channels through social media posts (Lobb, Mock & Hutchinson 2012). Research studies also confirmed that the amount of social media posts about the earthquake was a huge factor in increasing the financial assistance in terms of donation and support, which facilitated the disaster response efforts to a great extent (Houston et al. 2015).

Another use of disaster data in social media is being a source of mental, emotional and social support. Posts related to mental health could be connected with concerned health organizations for instance. Social media can also help connect survivors, affected individuals, and other concerned people to support each other through conversations or support groups (Taylor, M. et al. 2012). It also gives people the chance to express how they feel, post their good wishes and memorialize victims.

2.2.2. Twitter during disasters

Twitter is the most popular social media platform used for data analytics. It is an open and extensive platform that allows us to see how people from all over the world from different locations are discussing specific topics by searching for certain hashtags or keywords. It is also a more general social media platform with a targeted purpose such as LinkedIn for work-related networking. It has 353 million monthly active users with over 500 million tweets daily (Twitter, Inc. 2021; Dean 2021) which shows the high degree of user engagement and

information availability. In addition, tweets in specific time ranges or in specific languages can also be retrieved with their API. They allow a great level of flexibility in retrieving tweets with the required custom search parameters. Since Twitter allows for extremely fast propagation and circulation of many several types of information, this makes it a great source for data for studies and research (Simon et al. 2015). As discussed in the earlier section, it was seen how data retrieved from Twitter was important in handling multiple crisis and disaster situations.

Before a crisis event happens, it is helpful to have the public prepared to the greatest possible extent. Twitter is useful in this regard since it provides emergency services with the means to communicate with the public and convey essential information, conveys guidelines of how to act in a certain disaster event, and keep people updated about locations/progress of the undergoing disaster such as a hurricane, flood, etc.... (Martínez et al. 2018; Pogrebnyakov et al. 2018)

Information regarding evacuations or rescue efforts can also be communicated through official Twitter accounts for governmental organizations and emergency organizations. Information can also be provided to avoid specific areas with high risk, provide situational awareness, lessen panic, request for help, better coordinate rescue operations and organize donations for relief efforts (Avvenuti et al. 2016; Reuter, Hughes & Kaufhold 2018).

Tweets in disaster situations can also help organizations at a greater scale. For instance, after the 2011 earthquake in Japan, a tweet targeted at the American ambassador involved in the rescue efforts at the time, requested immediate assistance to transport critical patients from some hospital by air. The tweet was actually followed through and assistance troops were

sent for the patient transfer. There are several other similar cases where governmental organizations utilized information from tweets for emergency response (Girtelschmid et al. 2016; Alshareef & Grigoras 2017; Squicciarini, Tapia & Stehle 2017).

In another case, in Nepal in 2015, an immense amount of damage was inflicted as a result of the 7.8 magnitude earthquake. In a little bit over seventy hours after the first wave hit the country, over three thousand volunteers were already on the field trying to help the injured and survivors. Such a large assembly of volunteers was possible in such a short time because they were tagged in various tweets asking for help when the disaster first hit. The tweets were identified and categorized, allowing the volunteers to work hand in hand with the emergency workers to allow for much more efficient disaster management operations (Radianti, Hiltz & Labaka 2016).

Organizations are also venturing into utilizing tweets for man-made crisis events as well as natural disasters. For instance, in the Boston bombings in 2013, photos and information of the suspect were posted on Twitter and heavily retweeted, which helped the authorities to capture the suspect in a short period of time (Cassa et al. 2013).

Such portrayed examples show how useful Twitter is when dealing with crisis events and its popularity with both people and official entities alike. Users are now expecting authoritative entities to have knowledge of their tweets at such urgent times and started relying on that as a primary route for help requests. In a sample study, it was found that over 80% of the people expected help after they requested it on Twitter when tagging the official accounts for the organizations involved (Zoppi et al. 2016).

Another advantage of Twitter in disaster situations is contagion and cascade behaviors (Fabrega & Paredes 2013). In many situations, when a person keeps seeing multiple tweets and hashtags related to the same event, the probability that an individual will retweet or engage with those tweets significantly increases. This probability increases even more if the event in question is a disaster or a crisis event (Romero, Meeder & Kleinberg 2011; Fabrega & Paredes 2013). This behavior is often compared to crowds clapping at concerts, where people are more likely to join in the collective clapping even if they did not originally plan to (Budak, Agrawal & El Abbadi 2012). This sensation is quite helpful when people keep retweeting important or highly time-sensitive urgent tweets in a disaster situation which allows earlier discovery and increases the chances of those urgent requests to be seen by emergency response organizations that can help.

Many studies (Imran & Castillo 2015; Lai, She & Tao 2017; Aswani et al. 2018) have shown that the increased usage of Twitter during both natural and man-made disasters has greatly improved disaster management procedures while allowing for faster emergency response. From the examples seen, such data helps decrease injuries/deaths, saves people in need of help, reduces damages, finds missing people, shares emergency response guidelines, provides essential necessities like shelter/food/medications, circulates progress updates and encourages volunteering/donation initiatives as well.

With such a vast flow of data, not all disaster tweets are helpful to emergency services. Many tweets contain sympathy or emotional statements, old information, retweets of irrelevant information, and many more noisy tweets. These kinds of tweets are utterly useless to

emergency responders and might negatively impact the chances of finding the really important tweets with messages to deliver.

In summary, it can be seen that there is an enormous amount of data available in the tweets before, during, and after a disaster that is generated at a rapid pace. In order for twitter disaster data to be useful and actually be of assistance to emergency services, relevant information needs to be extracted first. Disaster tweets need to be analyzed if they are relevant or not, and classified into clear categories for more organized disaster management.

2.3. Machine Learning for Emergency Response

Understanding social media data to support emergency services faces many challenges, including analyzing short unstructured content, managing the overload of information, removing noisy data, filtering out irrelevant useless data, and many more. Several computational approaches and AI techniques for processing social media data have been proposed in the past couple of years to assist with effective disaster management. These approaches are designed to address multiple issues such as filtering relevant information, classification, handling overloading, categorization, and summarization (Imran & Castillo 2015; Alam et al. 2018).

Most of the AI methods employed for analyzing disaster social media data mainly use supervised or unsupervised techniques such as classification, clustering, and topic modelling. For tasks such as general text classification of social media textual data such as tweets, surveyed literature shows the implementation of both classic machine learning algorithms and deep learning-based approaches (Imran et al. 2013; Imran & Castillo 2015;

Alam et al. 2018). Classic algorithms used include Random Forest, Naïve Bayes, Logistic Regression, and Support Vector Machine (Burel & Alani 2018). The most commonly used deep learning-based techniques include Convolutional Neural Networks (CNN) and Long-Short-Term-Memory networks (LSTM) (Nguyen, Alam, et al. 2017).

2.3.1. Deep Learning

Deep learning models perform much better than the classic machine learning models when it comes to classification tasks, especially when used with pre-trained word embeddings (Goldberg 2015). There are several types of neural networks including long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM), recurrent neural networks (RNN), and convolutional neural networks (CNN) (Wu, Liu & Wang 2020).

CNNs are one of the most popular deep learning architectures used for classification tasks. The architecture of a CNN includes several convolution, pooling and dense fully connected layers (Goldberg 2015). In case of text classification, the convolutional model is targeted for feature extraction to learn salient features from the text input with the help of the word embeddings. So, in convolution layers, filters with various sizes are used to convolve the text matrix to detect features or patterns in the text (Zhou 2020). After that, max-pooling layers perform down-sampling on the input representation to reduce the dimensionality. Then, the fully connected layer interprets the extracted features for the prediction or classification result (Goldberg 2015; Zhou 2020).

When a CNN is dealing with image classification, the input image is convolved with a kernel to extract features. An activation function such as ReLU is applied to the convoluted values

to improve the non-linearity (Wu, Liu & Wang 2020). The pooling layer decreases the image size by retaining the important features and removing other areas from the image. This also plays a part in decreasing the cost of computation (Minaee et al. 2021). Max-pooling is the most popular pooling approach. The pooling matrix's size determines the degree of image reduction. For instance, a 2x2 pooling matrix will decrease the size of the image by 50% (Ribeiro, Singh & Guestrin 2016). The sequence of convolution and pooling layers helps in identifying the features. The pooled output from the stacked feature maps is fed to the next layer by flattening the maps. The last layers are dense fully connected layers. The last layer is responsible for the classification output (Goldberg 2015; Minaee et al. 2021).

(Nguyen, Al Mannai, et al. 2017) developed a CNN model to classify disaster tweets into useful or not useful. Disaster tweets are passed to the model's input layer and then transformed into a feature sequence. That is followed by a look-up layer that generates input vectors for the corresponding tokens, then passes them to the following sequence of convolution and max-pooling layers for learning the high-level feature representations (Nguyen, Al Mannai, et al. 2017). The convolution operation applied filters to generate the feature maps. They also implemented wide convolution so that those filters can cover the whole sentence with the boundary words as well. Zero padding was performed so that out-of-range vectors are taken as zeros (Nguyen, Al Mannai, et al. 2017). When the convolution operation is done, max-pooling is done for each of the feature maps. Their model also included a dense layer consisting of hidden nodes to handle variable lengths of sentences. The dense layer generates fixed size output vectors that are passed to the final output layer.

They also used the pre-trained Google embeddings (Mikolov et al. 2013) for initializing their models.

(Chaudhuri & Bose 2019) implemented a CNN model that would identify human body parts from images of wreckage and debris to help identify victims in need of rescue. All the images were preprocessed to confirm that they all have the same aspect ratio and size. The CNN model had four kinds of layers including a convolutional layer, max-pooling layer, rectified linear unit layer (ReLU) and fully connected layer (FC) (Chaudhuri & Bose 2019). The input layer holds the pixel values for the images, and the convolutional layers extract high-level features from the images. The pooling layer applies down-sampling across the spatial dimensions. The last layer responsible for the classification uses the softmax function (Chaudhuri & Bose 2019).

2.3.2. Disaster Social Media Text Classification

Classifying tweets in the context of disaster tweets aims to recognize whether a specific tweet is relevant or informative to emergency workers. With the enormous amount of data on Twitter, just classifying the relevancy is not enough, and further classification is required to assist emergency services properly. Therefore, there is a need to identify the category the disaster tweet belongs to such as asking for help, sharing information, medical assistance and so on (Vitale et al. 2012).

Several studies have performed the task of classifying disaster tweets into multiple classes. In (Caragea, Silvescu & Tapia 2016), SVM was used to classify disaster tweets regarding

the Haiti 2010 earthquake to multiple classes such as shortage of food, medical needs, emergencies and reached an F1-score of 0.59. Another study (Panagiotopoulos et al. 2016) used Naïve Bayes for two classification tasks achieving F1-scores ranging from 0.56 to 0.81. The first task is classifying tweets into informative, not-informative and personal. The second task is further classifying the informative tweets into multiple class categories including causalities, donations, information source, and caution.

Multiple machine learning approaches including logistic regression, KNN, SVM, decision trees, supervised latent Dirichlet allocation, and Naïve Bayes were used in (Ashktorab et al. 2014; Thom et al. 2015) to recognize disaster tweets that were specifically related to any victims or damages. The model with the best results was the logistic regression, achieving an F1-score of 0.65.

A deep learning model implementing a convolutional neural network (CNN) was proposed in (Nguyen, Al Mannai, et al. 2017) to classify disaster tweets to informative vs. non-informative classes as well. They also demonstrated how some data irrelevant to the disaster can be incorporated in the classifier's training process in the beginning phases of crisis events. Another study (Nguyen et al. 2016) also used deep learning models to classify disaster tweets to informative and non-informative classes, while also classifying the informative tweets further to other classes such as support, infrastructure damage and affected people.

Another study (Tatsubori et al. 2012) extracted disaster tweets related to two disasters: Hurricane Sandy in 2012 and the Joplin tornado in 2011 and attempted using a model to obtain relevant information from the informative disaster tweets, achieving a detection rate

of 49%. A similar study (Thom et al. 2015) also used disaster tweets related to hurricane Sandy in 2012 after manually labelling them to different classes first. Logistic regression was applied to perform the classification and an F1-score of an average 0.65 was achieved. Also dealing with disaster tweets on Hurricane Sandy, (Simon, Goldberg & Adini 2015) performed a study to classify tweets using a Naïve Bayes classifier that was also tested on another dataset for tweets regarding Boston Bombings in 2013 to experiment with both natural and man-made disasters.

Another known platform in the disaster tweets scene is AIDR, which stands for Artificial Intelligence for Disaster Response, a platform built by (Vieweg, Castillo & Imran 2014) to classify crisis-related tweets into specified classes as they happen. The creators utilized human resources together with machine learning models to label a dataset of disaster tweets and then trained models to automatically perform classification tasks on new tweets. In its initiation, AIDR was tested with disaster tweets data resulting from the 2013 earthquake in Pakistan to classify tweets to informative and not informative classes.

Another deep learning approach was implemented (Ragini, Anand & Bhaskar 2018) using CNN attempting to classify the crisis tweets into multiple labels. The authors also experimented with using some features in Twitter including keywords, mentions, retweets, and hashtags and tested their effect when incorporated with the tweets data while performing the classification task. They classified the tweets over seven classes including volunteering efforts, damages, awareness, specific information, not informative, causalities and sympathy with F1 scores ranging from 0.74 to 0.97.

Multiple models were compared in (Yu et al. 2019), where logistic regression, SVM, and CNN were used to perform classification of tweets related to the three hurricanes: Irma, Harvey and Sandy. The classification was done to several categories including Advice/Caution, Donations, Source of Information, Damage/Casualties and Resources/Infrastructures. The classification task was done with two settings using both data specific to a disaster and general data irrelevant to the disaster. F1-score achieved was between 0.30 to 0.79, with the CNN model having the best performance when compared to the other classic machine learning models.

2.3.3. Disaster Social Media Image Classification

Nowadays, the number of people who take pictures and share them on various social media platforms is very high, with almost everyone having a portable mobile phone. This allows individuals to share much more information regarding different activities in an easier and more expressive manner compared to just text. With the availability of such a huge number of images, there lies a lot of potential for analysis and extracting information. This potential is even higher when dealing with images related to disasters or crisis events since images would provide more details regarding the situation. For instance, a picture can show the exact state of infrastructure damage, a clearer look at surrounding environments, efficient recognition of victims or lost individuals and more accurate representations for emergency services to assist with a smoother disaster management operation (Ilyas 2014; Landwehr et al. 2016). Although images hold such a value, the amount of in-depth studies exploring the role of images on disaster social media in providing essential information is much less than

studies dealing with only text. The amount of studies in this field has been increasing over the past few years, given the valuable hidden information within disaster social media image data in assisting emergency services (Laylavi, Rajabifard & Kalantari 2017).

Multiple studies have reported the high importance of social media images produced during crisis events of disasters (Nguyen, Alam, et al. 2017; Alam et al. 2018; Burel & Alani 2018; Martínez-Rojas, Pardo-Ferreira & Rubio-Romero 2018a; Imran et al. 2020; Ray & Bala 2020). Some studies analyzed the metadata with the images and utilized the geotagging feature to retrieve locations of disaster-affected areas. The analysis of these images is experimented on using deep learning approaches. Since the amount of disaster images on social media is massive and quite noisy with irrelevant data, benefiting from them manually is not an option. Proper techniques that would filter out irrelevant data need to be developed. For instance, (Alam et al. 2018) created a pipeline to process disaster images retrieved from social media using deep learning approaches. The phases included gathering crisis-related pictures, filtering them and removing any irrelevant ones, and then classifying them.

Another study (Nguyen, Alam, et al. 2017) implemented deep learning techniques to detect irrelevant disaster images by incorporating transfer learning methodologies. They created a pipeline for recognizing irrelevant or repeated disaster pictures from social media platforms such as Twitter. They used CNN to classify the images to three damage degree classes: mild, severe and none. Datasets used included disaster images for Ruby typhoon, Ecuador earthquake and Nepal earthquake, achieving F1-scores between 0.66 to 0.85.

In (Boccignone et al. 2016), images related to fires in Australia were used in a locally grown dataset and classified into two classes: fire or not to recognize whether the image contained

a fire. The model achieved a high accuracy of around 85%. Another study (Chaudhuri & Bose 2019) implemented a CNN on a dataset of 514 images related to various earthquakes to identify human body parts from the wreckage pictures, achieving a high accuracy of a bit over 80%.

Some studies focus on identifying the degree of damage to damaged infrastructure, for instance, to help emergency services with swifter rescue operations. Unfortunately, the majority of such models are mostly trained with limited data lacking diversity, and thus are hard to generalize to many other kinds of natural or man-made disaster situations (Caragea, Silvescu & Tapia 2016; Panagiotopoulos et al. 2016). One of the main reasons for this limited performance is the lack of annotations for disaster images datasets. It is ideal to have clear annotations of various disasters or crisis events of different types on a large scale so that the analysis of images is done better allowing for more accurate assessment for damages. This shows the critical need for curating a large-scale disaster images dataset (Purohit et al. 2018).

There have also been increased improvements in facial recognition from images, which use social media images to identify missing people especially in the context of disasters. This would offer great help to emergency workers to better allocate resources to save the largest possible number of victims (Kushwaha et al. 2018; Kalliatakis et al. 2019).

2.3.4. Disaster Social Media: Multimodal Data

Given the various sources and characteristics of data, there is a need for machine learning algorithms that can combine these different aspects to make the most of the data and

maximize its usefulness (Baltrušaitis, Ahuja & Morency 2018). Multimodal learning addresses this point by using machine learning algorithms to build models that would learn from the different modalities. In the past few years, there has been gained interest in developing multimodal systems so that they can exploit both the text and images in the tweets to improve classification performance (Baltrušaitis, Ahuja & Morency 2019).

A huge percentage of social media posts are multimodal where text and image data are seen together. Multimodal data can sometimes contain supplemental information that can be quite beneficial to understand the broad view of a crisis event or a disaster with much more details when analyzed together. Several studies demonstrated that multimodal data analysis performs better with regards to classification, filtering relevant information and decreasing overload of information (Chen et al. 2013; Dewan et al. 2017; Martínez-Rojas, Pardo-Ferreira & Rubio-Romero 2018a). Moreover, there are several studies exploiting multimodal disaster tweets to assess damage levels and detection of events (Baltrušaitis, Ahuja & Morency 2018; Agarwal et al. 2020).

In (Jomaa, Rizk & Awad 2017), a multimodal model was developed where visual features were extracted from images and semantic features were extracted from text data for the feature vectors. The features were then used to train and develop a SVM classifier, resulting in improved performance compared to only-text or only-image models by around four percent. A similar study (Alqhtani, Luo & Regan 2015) attempting to detect when different events happen trained a KNN classifier on multimodal twitter data. They also extracted visual and semantic features from the images and text respectively, obtaining an

improvement in classification of around eight percent compared to any of the unimodal text-only or image-only models.

The affiliation between the text and images of the tweets was studied in (Chen et al. 2013) to classify relevant vs. irrelevant tweets. The study used the features extracted from the text and images together with other features such as retweets, number of followers, time of posting, etc. The model achieved an F1-score of around 71%, which was an improvement of about 6% compared to the classification by only textual data.

In (Rizk et al. 2019), a classifier for classifying multimodal disaster Twitter data was developed to classify to types either infrastructure damage (such as bridges or bridges) or nature damages (such as trees or forests). They used a home-grown dataset with tweets collected over three earthquakes and one flood disaster events. The study used visual features and characteristics from the tweet's picture and concatenated them with the semantic features extracted from the text. Their results showed that the multimodal model had better performance compared to when the model was only built using text-only or image-only data. A study (Mouzannar, Rizk & Awad 2018) used deep learning techniques to develop a multimodal system exploring damage detection and classifying disaster tweets to categories including floods, fires, infrastructure damage, nature damage, injured or dead individuals and no damage. They experimented with unimodal and multimodal models and also found that there is an improvement in the classification performance in the case of the multimodal model. A similar study (Gautam et al. 2019) also compared unimodal and multimodal models for crisis tweets, and used decision fusion to classify the text and images in those tweets to binary classes of informative vs. not informative.

A recent study (Ofli, Alam & Imran 2020) implements deep learning techniques with disaster tweets from the CrisisMMD dataset to develop classifiers for informativeness and humanitarian category. They experiment with unimodal models of only text and only images, together with multimodal data as well, also reaching the conclusion that multimodal models perform better than unimodal ones.

2.4. Summary

This chapter presented a comprehensive review of disasters, their presentation in social media and the utilization of such data with machine learning models to assist efficient disaster management. The work in this dissertation differs from previous work in several ways (Mouzannar, Rizk & Awad 2018; Gautam et al. 2019; Imran et al. 2020). For instance, (Mouzannar, Rizk & Awad 2018) is only focusing on classifying the environmental and human damages for disaster tweets from a home-grown dataset. Therefore, this causes some limitations to whether their findings can be generalized over all different disasters. However, this study is doing a comparative study between three representative disaster datasets in the crisis informatics research scene that are publicly available: CrisisMMD (Alam, Ofli & Imran 2018), CrisisNLP (Imran, Mitra & Castillo 2016), and CrisisLex26 (Olteanu, Vieweg & Castillo 2015). In addition, two classification tasks are implemented which are informativeness and humanitarian category. Although (Gautam et al. 2019) used a subset of CrisisMMD, they focused only on one classification task which is the informativeness. They also applied a decision fusion methodology for their multimodal deep learning models while this work applies feature fusion instead.

The recent study by (Imran et al. 2020) was considered a starting point for this study. They implemented deep learning models using disaster tweets from the CrisisMMD dataset for both classification tasks: informativeness and humanitarian category, and compared performance between unimodal and multimodal models. This dissertation starts with following their approach and using the CrisisMMD as a first dataset to compare results with theirs, then goes on to perform the mapping between the three representative crisis datasets used (CrisisMMD, CrisisNLP, and CrisisLex26) to have uniform classes for a more cohesive comparative analysis. Both classification tasks are performed for all the three datasets over unimodal and multimodal models with the classifications' performance results compared as well. Furthermore, a consolidated multimodal dataset from all three datasets is formed to serve as a new multimodal baseline dataset to achieve better results.

Chapter 3

Methodology

This chapter explains the methodology used in this study. It will present the system architecture, the three disaster datasets used in detail, data preprocessing, classification tasks for unimodal and multimodal data, and deep learning models used.

3.1. Overall System Architecture

The overview of the architecture and design of the multimodal system is presented in Figure 7. Deep learning approaches are used for all classification tasks over all modalities. A CNN network is used for processing the tweets' text. For text classification, a CNN is used with various filters and five hidden layers. For image classification, a VGG16 network is used where the fully connected layer of the network extracts high-level features of the disaster image. As for the multimodal classification, it is following a feature fusion approach. After both text and image classifications are done in parallel, two features will be acquired from both the text and image modalities. Then, these features will be input to form a shared representation of both modalities and later followed by a dense layer. Finally, softmax performs the prediction for the multimodal model.

Two classification tasks are performed for all three disaster datasets in three different settings:

- Classification tasks are informativeness and humanitarian category. Informativeness determines if either the disaster text or image is beneficial for emergency services, in which case it is labelled as “informative”, otherwise they are labelled as “not-

informative). Humanitarian Categories will determine the type or nature of the informative disaster tweet and classify it into a specific humanitarian category, as shown in the following respective sections of each dataset.

- Datasets are CrisisMMD, CrisisNLP, and CrisisLex26.
- Settings are unimodal text-only data, unimodal image-only data, and multimodal (data)

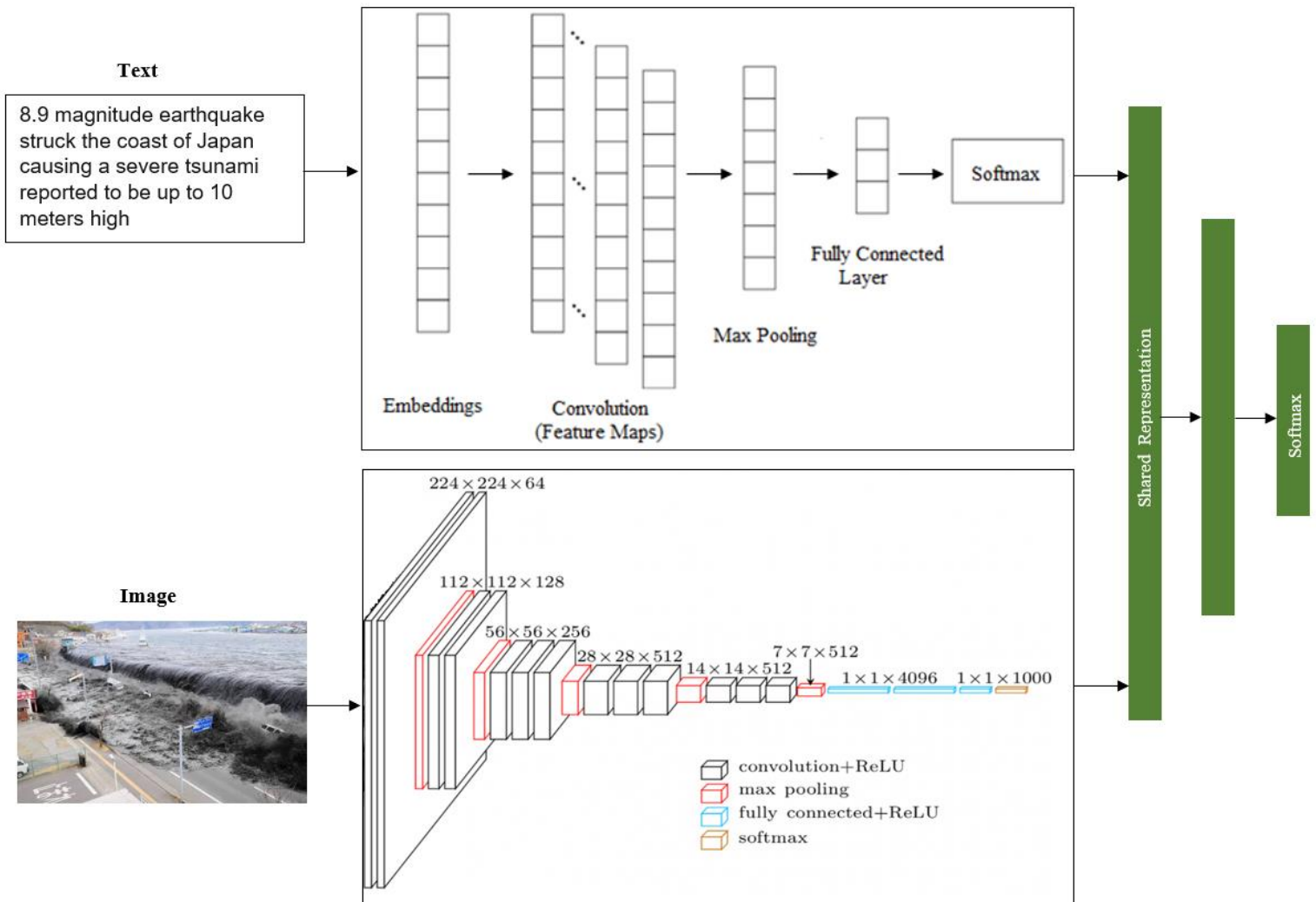






Figure 7: Overall Architecture of Multimodal Classification Approach

For some sample representations of the disaster tweets available in the datasets at hand, Table 2 shows some randomly selected multimodal tweets with their informativeness and humanitarian categories.

| | |
|--|--|
|  <p>Tweet Text: California wildfires kill 10, destroy 1,500 buildings Disaster: California Wildfires Informativeness: informative Humanitarian Category: infrastructure and utility damage</p> |  <p>Tweet Text: My car lost against Hurricane Maria when 60ft palm tree fell on top of it! Disaster: Hurricane Maria Informativeness: informative Humanitarian Category: vehicle damage</p> |
|  <p>Tweet Text: 1,500 bodies recovered; an equal number said to be injured Disaster: Nepal Earthquake Informativeness: informative Humanitarian Category: injured or dead people</p> |  <p>Tweet Text: Forget biking to work, today in Colorado we could have kayaked Disaster: Colorado Floods Informativeness: informative Humanitarian Category: affected individuals</p> |



| | |
|---|--|
|  <p>Tweet Text: Mexico City earthquake: 6 ways to help victims, from Airbnb to GoFundMe http://ift.tt/2w7FLW0 via FastCompany Disaster: Mexico Earthquakes Informativeness: informative Humanitarian Category: volunteering or donation</p> |  <p>Tweet Text: Hillary Clinton links climate change to recent wildfires, hurricanes in California speech Disaster: California Wildfires Informativeness: not-informative</p> |
|---|--|

Table 2: Sample Labelled Disaster Tweets

3.2. Datasets

This section presents the details of each of the three representative disaster datasets used, including their contents’ description, labelling, data retrieval, and properties breakdown.

3.2.1. First Dataset: CrisisMMD

CrisisMMD (Alam, Ofli & Imran 2018) is a multimodal dataset that consists of disaster tweets with images that were obtained throughout seven disasters that happened in 2017: Hurricanes (Irma, Harvey, Maria), Earthquakes (Mexico, Iraq/Iran), California wildfires, and Srilanka floods. The dataset has three annotations for three classification tasks: Informativeness (informative vs. not informative), eight classes for the humanitarian category, and three classes for damage severity (severe, mild, little to none). However, the

third classification task for damage severity is only applicable for images, so it is not taken into consideration in this study, where only the first two classification tasks are focused on. Table 3 shows the data distribution of the CrisisMMD dataset for the informativeness classes.

| Informativeness | |
|------------------------|--------------|
| Informative | 9374 |
| Not informative | 8708 |
| Total | 18082 |

Table 3: CrisisMMD Informativeness Distribution

Table 4 shows the data distribution of the CrisisMMD dataset for the humanitarian category classes.

| Humanitarian Category | |
|---|--------------|
| Affected individuals | 562 |
| Infrastructure and utility damage | 3624 |
| Injured or dead people | 110 |
| Missing or found people | 14 |
| Not humanitarian | 8708 |
| Other relevant information | 2529 |
| Rescue, volunteering or donation effort | 2231 |
| Vehicle damage | 304 |
| Total | 18082 |

Table 4: CrisisMMD Humanitarian Category Distribution

When CrisisMMD was in the creation stages, the text and image aspects of the disaster tweets were labelled separately, so there are some tweets where the text and image would have different labels (Alam, Ofli & Imran 2018). Therefore, the currently used dataset (Ofli, Alam & Imran 2020) is a subset of CrisisMMD where both text and image components of a tweet have the same label for either the informativeness or the humanitarian category classification tasks. It is also merging minority categories to ensure having a balanced distribution across all categories. This filtering resulted in having some inconsistencies with the label distribution. Therefore, they merged the similar categories (missing/found people and injured/people) into the (affected individuals) category. Moreover, the (vehicle damage) was merged with the (infrastructure/utility damage) class.

The same tweet can have a maximum of four images, so there are multiple records for the same text and different images. Therefore, it is essential to ensure that there are no repeated text tweets when splitting the dataset into training, development, and test datasets. The ratio used for splitting the dataset is 70% for the training dataset, 15% for the development dataset, and 15% for the test dataset. This is the same ratio implemented by (Ofli, Alam & Imran 2020) and will be used across all three datasets (CrisisMMD, CrisisNLP, CrisisLex26) to allow for a comparative analysis. The training dataset is for training the model, the development dataset is for tuning the parameters to avoid overfitting, and the test dataset is for evaluating the model.

Tables 5 and 6 show the data distribution after filtering and merging categories over the training, development and test datasets with corresponding labels for the informativeness and humanitarian category classification tasks respectively.

| | Train (70%) | Development (15%) | Test (15%) | Total |
|-----------------|------------------------|------------------------------|-----------------------|--------------|
| Informative | 6345 | 1056 | 1030 | 8431 |
| Non-informative | 3256 | 517 | 504 | 4277 |
| Total | 9601 | 1573 | 1534 | 12708 |

Table 5: Informativeness Data Split over CrisisMMD Subset

| | Train (70%) | Development (15%) | Test (15%) | Total |
|-------------------------------------|------------------------|------------------------------|-----------------------|--------------|
| Rescue/volunteering/donation effort | 912 | 149 | 126 | 1187 |
| Affected Individuals | 71 | 9 | 9 | 89 |
| Infrastructure/utility damage | 612 | 80 | 81 | 773 |
| Other relevant information | 1279 | 239 | 235 | 1753 |
| Not humanitarian | 3252 | 521 | 504 | 4277 |
| Total | 6126 | 998 | 955 | 8079 |

Table 6: Humanitarian Category Data Split over CrisisMMD Subset

3.2.2. Second Dataset: CrisisNLP

CrisisNLP is another representative disaster dataset in the Crisis Informatics field. It originally consists of around fifty million disaster tweets obtained during nineteen different natural and man-made disasters between 2013-2015: five earthquakes, three typhoons, one volcano, two floods, two wars, two biological, three landslides, and one airline accident.

Table 7 shows the data distribution for each of the disasters.

| Crisis Type | Crisis Name | Country | Language | Number of Tweets | Year |
|--------------------|---|----------------|-----------------|-------------------------|-------------|
| Earthquake | Nepal Earthquake | Nepal | English | 4,223,937 | 2015 |
| Earthquake | Terremoto Chile | Chile | Spanish | 842,209 | 2014 |
| Earthquake | Chile Earthquake | Chile | English | 368,630 | 2014 |
| Earthquake | California Earthquake | USA | English | 254,525 | 2014 |
| Earthquake | Pakistan Earthquake | Pakistan | English | 156,905 | 2013 |
| Typhoon | Cyclone PAM | Vanuatu | English | 490,402 | 2015 |
| Typhoon | Typhoon Hagupit | Philippines | English | 625,976 | 2014 |
| Typhoon | Hurricane Odile | Mexico | English | 62,058 | 2014 |
| Volcano | Iceland Volcano | Iceland | English | 83,470 | 2014 |
| Floods | Pakistan Floods | Pakistan | English | 1,236,610 | 2014 |
| Floods | India Floods | India | English | 5,259,681 | 2014 |
| War & Conflicts | Palestine Conflict | Palestine | English | 27,770,276 | 2014 |
| War & Conflicts | Peshawar School Attack | Pakistan | English | 1,135,655 | 2014 |
| Biological | Middle East Respiratory Syndrome (MERS) | Worldwide | English | 215,370 | 2014 |
| Biological | Ebola Virus Outbreak | Worldwide | English | 5,107,139 | 2014 |
| Landslide | Landslides worldwide | Worldwide | English | 382,626 | 2014 |
| Landslide | Landslides worldwide | Worldwide | French | 17,329 | 2015 |
| Landslide | Landslides worldwide | Worldwide | Spanish | 75,244 | 2015 |
| Airline Accident | Flight MH370 | Malaysia | English | 4,507,157 | 2014 |

Table 7: CrisisNLP Disaster Collections Distribution

The data was annotated by volunteers and paid workers. The CSV files provided included tweet ids and label for each separate disaster in a separate directory. The original CSV files were organized as follows:

Data labelled by volunteers:

1. 2014 California Earthquake
2. 2014 Chile Earthquake (Spanish tweets)
3. 2014 Chile Earthquake (English tweets)
4. 2014 Hurricane Odile in Mexico
5. 2014 Iceland Volcano
6. 2014 Malaysia Airline MH370 Accident
7. 2014 Middle East Respiratory Syndrome
8. 2014 Philippines Typhoon Hagupit
9. 2015 Cyclone Pam in Vanuatu
10. 2015 Nepal Earthquake
11. 2014 Worldwide Landslides (English tweets)
12. 2015 Worldwide Landslides (French tweets)
13. 2015 Worldwide Landslides (Spanish tweets)
14. 2014 Malaysia Airline Accident for Flight MH370

Data labelled by crowdsourcing:

1. 2013 Pakistan Earthquake
2. 2014 California Earthquake

3. 2014 Chile Earthquake (Spanish tweets)
4. 2014 Chile Earthquake (English tweets)
5. 2014 Ebola Virus
6. 2014 Hurricane Odile in Mexico
7. 2014 India Floods
8. 2014 Middle East Respiratory Syndrome
9. 2014 Pakistan Floods
10. 2014 Philippines Typhoon Hagupit
11. 2015 Cyclone Pam in Vanuatu
12. 2015 Nepal Earthquake

Since only English tweets are considered in this study, all tweets in languages other than English will be ignored. So, three directories are removed from the data labelled by volunteers, and one directory is removed from the data labelled by paid workers. Therefore, now the available data directories are:

Data labelled by volunteers:

1. 2014 California Earthquake
2. 2014 Chile Earthquake
3. 2014 Hurricane Odile in Mexico
4. 2014 Iceland Volcano
5. 2014 Malaysia Airline MH370 Accident
6. 2014 Middle East Respiratory Syndrome
7. 2014 Philippines Typhoon Hagupit

8. 2015 Cyclone Pam in Vanuatu
9. 2015 Nepal Earthquake
10. 2014 Worldwide Landslides
11. 2014 Malaysia Airline Accident for Flight MH370

Data labelled by crowdsourcing:

1. 2013 Pakistan Earthquake
2. 2014 California Earthquake
3. 2014 Chile Earthquake
4. 2014 Ebola Virus
5. 2014 Hurricane Odile in Mexico
6. 2014 India Floods
7. 2014 Middle East Respiratory Syndrome
8. 2014 Pakistan Floods
9. 2014 Philippines Typhoon Hagupit
10. 2015 Cyclone Pam in Vanuatu
11. 2015 Nepal Earthquake

Not all disaster tweets were labelled following the same labels. Some of the tweets were labelled using completely different categories than those followed in this study for the informativeness and humanitarian category tasks. The disasters that did not have the required labelling approach included:

1. 2014 Iceland Volcano

2. 2014 Malaysia Airline Accident for Flight MH370
3. 2014 Middle East Respiratory Syndrome
4. 2014 Worldwide Landslides
5. 2014 Ebola Virus

For instance, for the Ebola and Middle East Respiratory Syndrome tweets, the labels were neither informative vs. not-informative nor the humanitarian category. They were: Disease signs/symptoms, disease transmission, disease prevention, disease treatment, death reports, affected people, other useful information, and irrelevant. Therefore, those five disaster directories were not included as well. Now, the available disaster directories are:

Data labelled by volunteers:

1. 2014 California Earthquake
2. 2014 Chile Earthquake
3. 2014 Hurricane Odile in Mexico
4. 2014 Philippines Typhoon Hagupit
5. 2015 Cyclone Pam in Vanuatu
6. 2015 Nepal Earthquake

Data labelled by crowdsourcing:

1. 2013 Pakistan Earthquake
2. 2014 California Earthquake
3. 2014 Chile Earthquake
4. 2014 Hurricane Odile in Mexico

5. 2014 India Floods
6. 2014 Pakistan Floods
7. 2014 Philippines Typhoon Hagupit
8. 2015 Cyclone Pam in Vanuatu
9. 2015 Nepal Earthquake

There are disaster directories present in both the data labelled by volunteers and data labelled by paid workers. These disasters are:

1. 2014 California Earthquake
2. 2014 Chile Earthquake
3. 2014 Hurricane Odile in Mexico
4. 2014 Philippines Typhoon Hagupit
5. 2015 Cyclone Pam in Vanuatu
6. 2015 Nepal Earthquake

Therefore, those will be merged into one directory for each disaster to have a unified directory for the corresponding disaster tweets. Now, there are final nine disasters available after this filtering, which are:

1. 2013 Pakistan Earthquake
2. 2014 California Earthquake
3. 2014 Chile Earthquake
4. 2014 Hurricane Odile in Mexico
5. 2014 India Floods

6. 2014 Pakistan Floods
7. 2014 Philippines Typhoon Hagupit
8. 2015 Cyclone Pam in Vanuatu
9. 2015 Nepal Earthquake

Another issue is that the humanitarian categories used across all the data directories for those disasters are not consistent. So, for instance, there were 13 humanitarian categories for the California Earthquake tweets labelled by the volunteers:

1. Injured or dead people
2. Missing, trapped, or found people
3. Displaced people
4. Infrastructure and utilities
5. Shelter and supplies
6. Money
7. Volunteer or professional services
8. Animal management
9. Caution and advice
10. Personal updates
11. Sympathy and emotional support
12. Other relevant information
13. Not related or irrelevant

On the other hand, there were nine categories for the California Earthquake tweets labelled by the paid workers:

- 1- Injured or dead people
- 2- Missing, trapped, or found people
- 3- Displaced people and evacuations
- 4- Infrastructure and utilities damage
- 5- Donation needs or offers or volunteering services
- 6- Caution and advice
- 7- Sympathy and emotional support
- 8- Other useful information
- 9- Not related or irrelevant

So, before merging the tweets for both directories into one file for the California Earthquake disaster, the label inconsistencies need to be resolved. Semantically similar categories were merged to reduce the number of extra categories available. For instance, in the case mentioned above, the categories (Shelter and supplies, Money, Volunteer or professional services) were all mapped to the (Volunteering/Donation efforts) category. “Animal management” is mapped to “other relevant information”.

Since the study is a comparative analysis across all three datasets (CrisisMMD, CrisisNLP and CrisisLex26), consistency is essential when dealing with the data splits and classification classes to obtain a clearer uniform view for the comparison of the models’ performance. As shown in Table 6, CrisisMMD had five classes for the humanitarian category. Therefore, the categories in the CrisisNLP will be mapped to these categories in CrisisMMD as well. The mapping is as follows:

- Injured or dead people → Injured or dead people → Affected individuals

- Missing, trapped, or found people → Missing or found people → Affected individuals
- Displaced people and evacuations → Affected individuals
- Infrastructure and utilities damage → Infrastructure/utility damage
- Donation needs or offers or volunteering services → Rescue/volunteering/donation effort
- Caution and advice → Other relevant information
- Sympathy and emotional support → Other relevant information
- Other useful information → Other relevant information
- Not related or irrelevant → Not humanitarian

Table 8 shows the distribution of labelled tweets per disaster type for all nine disasters included.

| Disaster Name | Labelled by paid workers | Labelled by volunteers | Total |
|-----------------------------|---------------------------------|-------------------------------|--------------|
| California Earthquake | 1701 | 183 | 1884 |
| Chile Earthquake | 1932 | 440 | 2372 |
| Cyclone Pam | 2004 | 600 | 2604 |
| Hurricane Odile Mexico | 1262 | 183 | 1445 |
| India Floods | 1820 | - | 1820 |
| Nepal Earthquake | 3003 | 9471 | 12474 |
| Pakistan Earthquake | 1881 | - | 1881 |
| Pakistan Floods | 1769 | - | 1769 |
| Philippines Typhoon Hagupit | 2010 | 9675 | 11685 |
| Total | 17382 | 20552 | 37934 |

Table 8: CrisisNLP Disaster Distribution for Labelled Tweets

The first dataset, CrisisMMD, had disaster tweets that had both text and image components during retrieval. However, the CrisisNLP dataset only has the text modality for the disaster tweets. In order to utilize it in a multimodal setting, tweets containing both text and images need to be identified, and the images need to be obtained as well.

So, first of all, a CSV file containing all the tweet ids for the 17382 disaster tweets labelled by paid workers was fed to a Python script to retrieve the tweet text and image URL for each. Twitter Developer API was used to make the connection, and secret keys were created to gain approval to retrieve the tweets. Unfortunately, since the tweets in this dataset are rather old (2013-2015), many of the users' accounts that posted the tweets were suspended, so those tweets could not be retrieved. So out of the 17382 tweets, only 10746 were still available. Then, since we are only interested in the tweets with both text and image modalities, further filtering was done to check which tweets returned a non-null value for the `image_url` field in the json response retrieved from the Twitter API. As a result, we were left with only 1020 tweets with both text and image modalities available.

Similarly, the 20552 disaster tweets labelled by volunteers were fed to the retrieval script. Only 13395 tweets were still available, while only 1511 tweets had both the text and image modalities. Therefore, after merging the tweets from the ones labelled by the volunteers and paid workers, we are left with a total of 2531 tweets.

Table 9 shows the distribution of tweets per disaster type for all nine disasters after this filtering has been done.

| Disaster | No. of Tweets |
|-----------------------------|----------------------|
| California Earthquake | 125 |
| Chile Earthquake | 82 |
| Cyclone Pam | 194 |
| Hurricane Odile Mexico | 121 |
| India Floods | 60 |
| Nepal Earthquake | 864 |
| Pakistan Earthquake | 52 |
| Pakistan Floods | 102 |
| Philippines Typhoon Hagupit | 931 |
| Total | 2531 |

Table 9: CrisisNLP Disaster Distribution after Filtering

Tables 10 and 11 show the data distribution of the CrisisNLP dataset for the informativeness and humanitarian category classes respectively.

| Informativeness | |
|------------------------|-------------|
| Informative | 1643 |
| Not informative | 888 |
| Total | 2531 |

Table 10: CrisisNLP Informativeness Distribution

| Humanitarian Category | |
|-------------------------------------|-------------|
| Affected individuals | 44 |
| Infrastructure and utility damage | 226 |
| Injured or dead people | 81 |
| Missing or found people | 32 |
| Not humanitarian | 888 |
| Other relevant information | 925 |
| Rescue/volunteering/donation effort | 335 |
| Total | 2531 |

Table 11: CrisisNLP Humanitarian Category Distribution

Tables 12 and 13 show the data distribution after filtering and merging categories over the training, development and test datasets with corresponding labels for the informativeness and humanitarian category classification tasks respectively.

| | Train (70%) | Development (15%) | Test (15%) | Total |
|-----------------|------------------------|------------------------------|-----------------------|--------------|
| Informative | 1155 | 237 | 251 | 1643 |
| Non-informative | 616 | 143 | 129 | 888 |
| Total | 1771 | 380 | 380 | 2531 |

Table 12: Informativeness Data Split over CrisisNLP Dataset

| | Train (70%) | Development (15%) | Test (15%) | Total |
|-------------------------------------|------------------------|------------------------------|-----------------------|--------------|
| Rescue/volunteering/donation effort | 232 | 58 | 45 | 335 |
| Affected Individuals | 99 | 25 | 33 | 157 |
| Infrastructure/utility damage | 151 | 43 | 32 | 226 |
| Other relevant information | 651 | 134 | 140 | 925 |
| Not humanitarian | 638 | 120 | 130 | 888 |
| Total | 1771 | 380 | 380 | 2531 |

Table 13: Humanitarian Category Data Split over CrisisNLP Dataset

3.2.3. Third Dataset: CrisisLex26

CrisisLex26 (Olteanu, Vieweg & Castillo 2015) is a dataset collection that consists of disaster tweets collected during twenty-six large disasters in 2012 and 2013. There are about 1000 tweets labelled by informativeness, source, and information type (CrisisLex: Crisis Collections n.d.).

The original unlabeled dataset had around 250 thousand tweets posted throughout the 26 crisis events, where most disasters had around two to four thousand tweets. Crowdsourcing workers were employed for the labelling task. Around 1000 tweets from each disaster collection were selected. The tweets were labelled by informativeness (informative vs. non-informative), source (government, etc.), and information type (advice, infrastructure damage, etc.) (CrisisLex: Crisis Collections n.d.; Olteanu, Vieweg & Castillo 2015). The dataset is provided as CSV files that contain the tweet ids and the labels given to the corresponding tweets. The dataset is available on the GitHub page (CrisisLex GitHub Repository n.d.).

The following are the disasters available in the dataset:

1. 2012 Colorado wildfires
2. 2012 Costa Rica earthquake
3. 2012 Guatemala earthquake
4. 2012 Italy earthquakes
5. 2012 Philippines floods
6. 2012 Typhoon Pablo
7. 2012 Venezuela refinery
8. 2013 Alberta floods
9. 2013 Australia bushfire
10. 2013 Bohol earthquake
11. 2013 Boston bombings
12. 2013 Brazil nightclub fire
13. 2013 Colorado floods
14. 2013 Glasgow helicopter crash
15. 2013 LA airport shootings
16. 2013 Lac Megantic train crash
17. 2013 Manila floods
18. 2013 NY train crash
19. 2013 Queensland floods
20. 2013 Russia meteor
21. 2013 Sardinia floods

22. 2013 Savar building collapse
23. 2013 Singapore haze
24. 2013 Spain train crash
25. 2013 Typhoon Yolanda
26. 2013 West Texas explosion

As mentioned in the previous section, consistency for the humanitarian categories is essential for a clearer comparison of models' performance at the end. So, mapping for the categories is done to have uniform classes throughout all datasets. As shown in Table 11, CrisisLex26 has seven classes for the humanitarian category. Therefore, the categories in the CrisisLex26 will be mapped to the categories in CrisisMMD as well as CrisisNLP. The mapping is as follows:

- Affected individuals → Affected individuals
- Infrastructure and utilities → Infrastructure and utility damage
- Donations and volunteering → Rescue, volunteering or donation effort
- Caution and advice → Other relevant information
- Sympathy and support → Other relevant information
- Other Useful information → Other relevant information
- Not applicable → Not humanitarian

CrisisLex26 dataset only has the text modality for the disaster tweets like CrisisNLP, unlike the CrisisMMD dataset that has both text and image modalities. Therefore, as done in the previous section for CrisisNLP, only tweets containing both text and images will be identified, and their corresponding images will be retrieved.

The dataset has separate CSV files for all 26 disasters. All these CSV files will be merged in a single CSV file representing all the labelled disaster tweets available. The tweets retrieval procedure is the same one followed in the previous section for the CrisisNLP dataset. The merged CSV file is fed to a Python script that retrieves the tweet text and image URL for each tweet. Then, the Twitter Developer API is used to make the connection with the creation of secret keys for approval to retrieve the tweets.

Once again, we face the issue of the tweets being rather old since they are obtained during 2012 and 2013, so even older than the tweets in CrisisNLP. Therefore, several user accounts who posted the tweets were suspended, so their tweets could not be retrieved. The original CSV file had 22271 tweets. After running the retrieval script, and further filtering to ignore non-available tweets of suspended accounts or text-only tweets without images, we are left with only 975 tweets with both text and image modalities available. Another thing to note is that among those 975 tweets, 133 tweets were only labelled for the informativeness task and were not labelled for the humanitarian task. Therefore, a total of 975 tweets are used for the informativeness classification task, and a total of 842 tweets are used for the humanitarian classification task.

Table 14 shows the distribution of tweets per disaster type for all 26 disasters included.

| Disaster Name | No. of Tweets |
|-------------------------------|----------------------|
| 2012 Colorado wildfires | 20 |
| 2012 Costa Rica earthquake | 13 |
| 2012 Guatemala earthquake | 19 |
| 2012 Italy earthquakes | 16 |
| 2012 Philippines floods | 13 |
| 2012 Typhoon Pablo | 33 |
| 2012 Venezuela refinery | 20 |
| 2013 Alberta floods | 87 |
| 2013 Australia bushfire | 72 |
| 2013 Bohol earthquake | 28 |
| 2013 Boston bombings | 24 |
| 2013 Brazil nightclub fire | 7 |
| 2013 Colorado floods | 44 |
| 2013 Glasgow helicopter crash | 69 |
| 2013 LA airport shootings | 29 |
| 2013 Lac Megantic train crash | 33 |
| 2013 Manila floods | 48 |
| 2013 NY train crash | 58 |
| 2013 Queensland floods | 77 |
| 2013 Russia meteor | 21 |
| 2013 Sardinia floods | 75 |
| 2013 Savar building collapse | 16 |
| 2013 Singapore haze | 45 |
| 2013 Spain train crash | 21 |
| 2013 Typhoon Yolanda | 63 |
| 2013 West Texas explosion | 24 |
| Total | 975 |

Table 14: CrisisLex26 Disaster Distribution for Labelled Tweets

Tables 15 and 16 show the data distribution of the CrisisLex26 dataset for the informativeness and humanitarian category classes respectively.

| Informativeness | |
|------------------------|------------|
| Informative | 617 |
| Not informative | 358 |
| Total | 975 |

Table 15: CrisisLex26 Informativeness Distribution

| Humanitarian Category | |
|-------------------------------------|------------|
| Affected individuals | 106 |
| Infrastructure and utility damage | 113 |
| Not humanitarian | 46 |
| Other relevant information | 478 |
| Rescue/volunteering/donation effort | 96 |
| Total | 842 |

Table 16: CrisisLex26 Humanitarian Category Distribution

Tables 17 and 18 show the data distribution after filtering and merging categories over the training, development and test datasets with corresponding labels for the informativeness and humanitarian category classification tasks respectively.

| | Train (70%) | Development (15%) | Test (15%) | Total |
|-----------------|------------------------|------------------------------|-----------------------|--------------|
| Informative | 426 | 97 | 94 | 617 |
| Non-informative | 256 | 49 | 53 | 358 |
| Total | 682 | 146 | 147 | 975 |

Table 17: Informativeness Data Split over CrisisLex26 Dataset

| | Train (70%) | Development (15%) | Test (15%) | Total |
|-------------------------------------|------------------------|------------------------------|-----------------------|--------------|
| Rescue/volunteering/donation effort | 68 | 15 | 13 | 96 |
| Affected Individuals | 74 | 17 | 15 | 106 |
| Infrastructure/utility damage | 70 | 20 | 23 | 113 |
| Other relevant information | 338 | 69 | 71 | 478 |
| Not humanitarian | 38 | 6 | 5 | 49 |
| Total | 588 | 127 | 127 | 842 |

Table 18: Humanitarian Category Data Split over CrisisLex26 Dataset

3.3. Data Preprocessing

The tweet texts in all the datasets are quite noisy. They have several emoticons, symbols, and invisible characters as well. Therefore, when preprocessing them, any stop words will be removed. Also, numbers, hashtags, non-ASCII characters, and URLs will be removed. Punctuation marks will also be replaced with white spaces instead. So, symbols including “%”, “&”, “@”, “#”, “!”, and others are removed from the tweets. Also, all text in the tweets is converted to lower case. Redundant tweets will be filtered out as well. The preprocessing steps are following the approach followed in (Ofli, Alam & Imran 2020).

As for the images obtained from the disaster tweets, typical preprocessing steps are followed. The pixels of each image between zero and one are scaled, followed by normalization of each channel with respect to the ImageNet dataset (Ofli, Alam & Imran 2020) when scaling the images. Then, when normalizing, the matrix of pixels of each image would be divided by the maximum value, which is 225 in this case. Then, the normalized matrix is used by the system developed for training and testing the machine learning model (Deng et al. 2010; Kumar et al. 2020)

3.4. Classification Approach

The approach for classification tasks for text, image, and multimodal modalities implemented in this work is following the one proposed in (Ofli, Alam & Imran 2020). Since this is a comparative study, the same values and approach implemented by the authors in (Ofli, Alam & Imran 2020) are used when handling all the models for all the datasets used in this work to achieve consistent comparable results.

3.4.1. Text Modality

Convolutional Neural Networks (CNN) are used in text classification because they are proven to have better performance when applied to disaster tweets' classification tasks (Nguyen, Al Mannai, et al. 2017). Therefore, a CNN that has five hidden layers is created. In the beginning, to handle the network's input, the disaster tweets are zero-padded to obtain an equal length. Considering that each row would represent a word in the extracted disaster tweet from a pre-trained word2vec model, the tweets are converted to a word-level matrix.

The pre-training of the word2vec model is discussed in detail in (Alam, Joty & Imran 2018). (Imran, Mitra & Castillo 2016) developed the first largest word2vec word embeddings specifically for Crisis informatics research, trained with fifty-two million disaster tweets. The Continuous Bag of Words (CBOW) approach from (Mikolov et al. 2013) is used to train this word2vec model. The CBOW is implemented on a very large dataset of 364 million disaster tweets and around three billion words, vector dimensions of 300, $k = 5$ negative samples, and a context window size of five (Alam, Joty & Imran 2018; Ofli, Alam & Imran 2020)

Now that the input to the network is prepared, it goes through several layers that include the convolutional layer and the max pooling layer. Then, we get a high-level feature representation that is fixed in size for each disaster tweet. Then, the feature vectors obtained are passed through the fully connected hidden layers then the output layer at the end. Rectified linear units (ReLU) (Krizhevsky, Sutskever & Hinton 2012) are used as the activation function in the convolutional and fully connected layers. The softmax activation function is used for the output layer.

The Adam optimizer (Zeiler 2012) is used for training the CNN models. As for the optimization of the classification loss with respect to the development subset, a learning rate of 0.01 was used. A maximum of 50 was set to the number of epochs. Also, a 0.02 dropout rate is used to combat overfitting (Srivastava et al. 2014).

Early stopping was also implemented with a patience of 10. In addition, the following filters are used: 100 filters with a window size of 2, 150 filters with a window size of 3, and 200 filters with a window size of 4. The pooling length used is the same as the size of the filter

window. Moreover, since batch normalization is proved to be successful (Ioffe & Szegedy 2015), it is applied here as well.

3.4.2. Image Modality

VGG-16 is a deep convolutional neural network trained on a subset of the ImageNet dataset (Deng et al. 2010). ImageNet is a collection of over 14 million high-resolution images labelled belonging to around 22 thousand categories. VGG-16 was proposed by (Simonyan & Zisserman 2015) and was trained on millions of images to classify 1000 various categories (Deng et al. 2010). It has previously achieved a high classification accuracy of 92.7% in the ImageNet Classification Challenge in 2014 (Ioffe & Szegedy 2015). As shown in Figure 8, the VGG-16 model has thirteen convolutional layers, three fully connected layers, and finally an output layer of one thousand nodes.

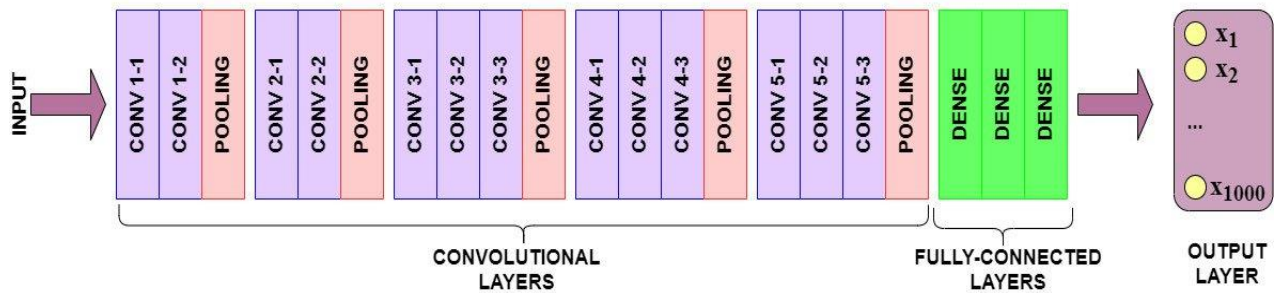


Figure 8: VGG-16 Model Architecture

The input to the VGG-16 network is an image of a fixed size “224 x 224 x 3”, then the convolution operation will be applied using a “3x3” filter. The five max pooling layers will

perform spatial pooling (Simonyan & Zisserman 2015). The third fully connected layer outputs one thousand channels for the thousand classes, then its output is passed to the final softmax layer for normalization of the classification vector (Krizhevsky, Sutskever & Hinton 2012). Further details with thorough descriptions of the mechanism of the VGG-16 network are shown (Deng et al. 2010; Simonyan & Zisserman 2015).

The architecture of VGG-16 is considered the optimal choice when extracting features from images. That has been proved in several studies employing image classification (Simonyan & Zisserman 2015; Nguyen, Al Mannai, et al. 2017; Nguyen, Alam, et al. 2017; Nguyen, Ofli, et al. 2017). Therefore, it is used in this study as well for the image modality classification of the disaster tweets' images.

When it comes to image classification, the deep learning model needs to learn how to recognize generic features like the edges for instance, then go on to detect more complicated features (Hussain, Bird & Faria 2019). To perform such an operation in a real-life setting, it would require many millions of pictures and very long training periods to get a decent performance. That is why using a pre-trained model and retraining it on the dataset of interest would be a much easier and efficient approach (Krishna & Kumar Kalluri 2019). This process is called transfer learning. So, transfer learning is mainly about using a machine learning model that was previously trained on some ML task, then reuse it as an initial point to start a different task. The transfer learning approach is based on the idea of using existing weights from a pre-trained model. So, the weights from a VGG16 model pre-trained on ImageNet are used for the model's initialization. (Yosinski et al. 2014; Kaur & Gandhi 2020)

show that transfer learning approaches are effective when used for recognizing visuals; therefore, they are employed in this study when handling the image modality classification.

The fully connected layer in the VGG-16 network generates one thousand output labels. However, the softmax layer at the end of the VGG-16 network is adapted depending on the specific classification type (informativeness and humanitarian category classification tasks). So, the images are passed through the convolutional layers of the VGG16 network, generating a feature stack containing the visual features that were recognized (Hussain, Bird & Faria 2019). This three-dimensional stack of features needs to be flattened so that it can be used by other ML classifiers for prediction tasks (Kaur & Gandhi 2020), so the stack is flattened to a NumPy array of pixel data. The pixel values would then be scaled for the VGG16 model, and the feature map can be generated.

Rectified linear units (ReLU) (Krizhevsky, Sutskever & Hinton 2012) are used as the activation function in the convolutional and fully connected layers. As mentioned before, the softmax activation function is used for the output layer. The Adam optimizer (Zeiler 2012) is used for training the image models with a starting learning rate of 10^{-6} . When the accuracy on the development subset does not seem to improve after 100 epochs, this learning rate would be then reduced by a factor of 0.1. Early stopping is also applied here with 1000 as the maximum number of epochs.

3.4.3. Multimodal Classification

The architecture of the multimodal deep neural network used is shown in figure 8. A VGG-16 network is used for the image modality, and a CNN network is used for the text modality. Shared representations are formed from both the text and image modalities. Right before the formation of these shared representations, a hidden layer of size 1000 from each side is employed. The same size is used from both modalities to ensure having an equal share. The size of 1000 can be modified, but it will be used as 1000 here as followed in (Ofli, Alam & Imran 2020) for the comparative study.

Then, a hidden layer is added before the softmax output layer after concatenating both the text and image modalities. The Adam optimizer is used when training the multimodal model, employing a batch size of 32. Early stopping is also used here to combat overfitting. In addition, ReLU is used as an activation function. (Ofli, Alam & Imran 2020) did not perform any tuning for the hyper-parameters such as the filter size or the hidden layers' size, so no tuning was done in this experiment as well to allow for proper comparison of models' performance later on.

Chapter 4

Results and Discussion

As mentioned in the previous sections, six classification tasks were performed over all three representative datasets. For each disaster dataset from CrisisMMD, CrisisNLP, and CrisisLex26, two classification tasks (informativeness and humanitarian category) were performed for three models (unimodal text-only data, unimodal image-only data, and multimodal with both text and image data).

When performing these classification tasks, the performance of all systems is evaluated using the accuracy, precision, recall, and F1-score. The description of these performance evaluation metrics is detailed in the next section.

4.1. Performance Evaluation Metrics

All evaluation metrics are calculated using four parameters: TP (true positive), TN (true negative), FP (false positive), and FN (false negative).

- ***True positives (TP)***: These are the positive values predicted correctly. So, the actual positive value of the class matches the predicted value (Abdelhade, Soliman & Ibrahim 2018). For instance, if the disaster tweet was labelled by humans as informative, and the predicted label by the model is also informative.

- **True Negatives (TN):** These are the negative values predicted correctly. So, the actual negative value of the class matches the predicted value (Abdelhade, Soliman & Ibrahim 2018). For instance, if the disaster tweet was labelled by humans as not informative, and the predicted label by the model is also not informative.
- **False Negatives (FN):** When the true value is positive but the model predicts a negative value (Goutte & Gaussier 2005). For example, if the disaster tweet was labelled by humans as informative, and the model predicts its label as not informative.
- **False Positives (FP):** When the true value is negative but the model predicts a positive value (Goutte & Gaussier 2005). For example, if the disaster tweet was labelled by humans as not informative, and the model predicts its label as informative.

Accuracy is the ratio of correctly predicted classification labels, including TP and TN, to the total number of available data instances (Goutte & Gaussier 2005). Higher accuracy does not necessarily mean better performance. It is a better performance measure when the datasets at hand are symmetric with the number of false negatives and false positives are close (Jeni, Cohn & De La Torre 2013). In the case of the informativeness classification task, accuracy would be answering the question: How many disaster tweets did the model correctly classify (both TP and TN) out of all the tweets?

Accuracy is calculated using the following formula (Goutte & Gaussier 2005):

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Precision is the ratio of the predicted positive values compared to the total predicted positive values (Sokolova, Japkowicz & Szpakowicz 2006). So, precision would be answering the question: How many of the disaster tweets labelled by the model as informative are actually informative?

Precision is calculated using the following formula (Goutte & Gaussier 2005):

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

Recall is the ratio of positives correctly predicted by the model to all the available actual positives (Sokolova, Japkowicz & Szpakowicz 2006). So, recall is answering the question: Of all the disaster tweets that are informative, how many of those did the model correctly predict as informative?

Recall is calculated using the following formula (Goutte & Gaussier 2005):

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

F1 score is the weighted average of both recall and precision. It takes both FP and FN into consideration to have a balance between recall and precision (Jeni, Cohn & De La Torre 2013). F1 score is seen as a better measure than accuracy in some cases where there is an uneven class distribution (Abdelhade, Soliman & Ibrahim 2018).

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.2. Models' Performance for CrisisMMD

This study uses Python 3.7.10. Table 19 shows the Python machine learning libraries used with their corresponding versions.

| Python Library | Version |
|---|---------|
| Keras (<i>Keras</i> n.d.) | 2.5.0 |
| pandas (<i>Pandas</i> n.d.) | 1.1.5 |
| nltk (<i>Natural Language ToolKit (nltk)</i> n.d.) | 3.5 |
| gensim (<i>gensim</i> n.d.) | 3.8.3 |
| numpy (<i>NumPy</i> n.d.) | 1.19.5 |
| scikit-learn (<i>scikit-learn</i> n.d.) | 0.22.2 |

Table 19: Machine Learning Python Libraries used

Tables 20 and 21 show the results for the informativeness and humanitarian category classification tasks respectively for the CrisisMMD with the different modalities when the deep learning models were run in this study. The results are consistent with the previously published results in (Ofli, Alam & Imran 2020). Further details for the comparison between both works can be found in Appendix A.

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|---------------|--------------|----------|-----------|--------|----------|
| Unimodal | Text | 81.4 | 81.0 | 81.4 | 81.0 |
| Unimodal | Image | 83.6 | 83.4 | 83.6 | 83.5 |
| Multimodal | Text + Image | 84.5 | 84.3 | 84.0 | 84.3 |

Table 20: Results for informativeness classification task for CrisisMMD

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|---------------|--------------|----------|-----------|--------|----------|
| Unimodal | Text | 69.0 | 69.0 | 69.0 | 66.0 |
| Unimodal | Image | 77.4 | 77.0 | 77.4 | 76.8 |
| Multimodal | Text + Image | 77.5 | 76.8 | 77.5 | 76.9 |

Table 21: Results for humanitarian category classification task for CrisisMMD

It can be seen that for the unimodal image-only models have a better performance when compared to the unimodal text-only models in both the informativeness and humanitarian category classification tasks. For the informativeness task, the unimodal image-only image is around 2% higher on the average of the metrics' values when compared to the unimodal text-only model. For the humanitarian category classification task, the unimodal image-only model is a little bit over 6% on the metrics' values average compared to the only text unimodal model.

The image-only unimodal model is performing better than the text-only unimodal model. That could be due to the fact that the text in each tweet is limited to 140 characters, whereas the image has all its characteristics with no restrictions. Images performing better than text in such classification models is also reported in other studies that were using tweets as their source of data (Gupta et al. 2013; Imran et al. 2015; Kumar et al. 2020).

The results also show that the multimodal model has the best performance in both informativeness and humanitarian categories. So, disaster data using both text and images together performs better than text alone or images alone. These findings further confirm that

multimodal learning leads to an improvement in performance when compared to unimodal learning.

For the humanitarian category classification task, all models are found to have a lower performance compared to the informativeness classification task. This is because the informativeness classification task is simpler with only two class labels (informative vs not informative). On the other side, the humanitarian category classification is more complex with more classes as shown in Table 4.

Another aspect that could be improved is a more balanced distribution of tweets in each humanitarian category. As seen in tables 4, 11, and 16 for the number of tweets per each humanitarian category for CrisisMMD, CrisisNLP and CrisisLex26 respectively, the categories “not humanitarian” and “other relevant information” usually have higher corresponding tweets compared to the other categories.

4.3. Models’ Performance for CrisisNLP

Tables 22 and 23 show the results for the informativeness and humanitarian category respectively for the CrisisNLP with the different modalities in this study.

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|---------------|--------------|----------|-----------|--------|----------|
| Unimodal | Text | 71.8 | 71.7 | 72.0 | 71.8 |
| Unimodal | Image | 70.0 | 69.0 | 70.0 | 69.2 |
| Multimodal | Text + Image | 70.0 | 70.0 | 70.0 | 64.0 |

Table 22: Results for informativeness classification task for CrisisNLP

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|---------------|--------------|----------|-----------|--------|----------|
| Unimodal | Text | 49.2 | 48.0 | 49.2 | 46.4 |
| Unimodal | Image | 46.3 | 44.0 | 46.3 | 44.2 |
| Multimodal | Text + Image | 48.4 | 51.0 | 48.0 | 45.2 |

Table 23: Results for humanitarian category classification task for CrisisNLP

The results for the CrisisNLP dataset are lower than CrisisMMD for both the informativeness and humanitarian category classification tasks. For the informativeness task, the unimodal text-only model here is around 9% less than CrisisMMD. The unimodal image-only model is around 13.5% less than CrisisMMD. The multimodal model here is around 14% lower as an average of most metrics, while the F1 score is around 20% less than CrisisMMD.

The performance for the humanitarian category for all modalities is significantly less than the results of CrisisMMD. The unimodal text-only model here is about 20% less than CrisisMMD. The unimodal image-only model is around 30% less than CrisisMMD. The multimodal model is around 29% less than CrisisMMD.

The decrease in the performance for all modalities in CrisisNLP is due to the much smaller dataset size. 12708 records were available in CrisisMMD for the informativeness task, while only 2531 records were available here in CrisisNLP. Similarly, 8079 records were available in CrisisMMD for the humanitarian category, while only 2531 records were available here in CrisisNLP. With deep learning models being hungry for data, the larger the size of the dataset, the higher chance it gets to properly learn and train the model (Aggarwal 2018). The performance in the humanitarian category classification task was lower when compared to the informativeness task because of the difference in complexity between them. As

mentioned in the previous section, the informativeness task classifies disaster tweets into only two labels (informative vs not informative), while the humanitarian category has more classes to predict as shown in Table 4.

The same finding reached in the previous section with CrisisMMD is reflected here with CrisisNLP as well: Multimodal models perform better than their unimodal counterparts. This has been found to be true for both classification tasks in both the datasets so far.

To accommodate for the lower performance results for CrisisNLP, a consolidation approach is attempted. Since all datasets were restructured to have the same structure including the number of columns in their corresponding CSV files for consistency, this allows us to consider consolidation. Therefore, the data from CrisisMMD is added to the data for CrisisNLP to generate a consolidated dataset of a larger size so that the models have much more data to learn from. Then, the experiment is repeated where all the models are rerun again for both classification tasks. Tables 24 and 25 show the results for the informativeness and humanitarian category respectively for the CrisisNLP consolidated with CrisisMMD in with the different modalities.

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|----------------------|-----------------|-----------------|------------------|---------------|-----------------|
| Unimodal | Text | 81.9 | 81.5 | 81.9 | 81.5 |
| Unimodal | Image | 83.7 | 83.5 | 83.7 | 83.6 |
| Multimodal | Text + Image | 84.7 | 84.5 | 84.2 | 84.5 |

Table 24: Results for informativeness classification task for CrisisNLP+CrisiMMD

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|---------------|--------------|----------|-----------|--------|----------|
| Unimodal | Text | 70.1 | 70.1 | 70.1 | 67.1 |
| Unimodal | Image | 77.5 | 77.1 | 77.5 | 77.0 |
| Multimodal | Text + Image | 77.8 | 77.2 | 77.7 | 77.2 |

Table 25: Results for humanitarian category task for CrisisNLP+CrisisMMD

The results after the consolidation of CrisisMMD and CrisisNLP are better than results for any of the datasets on their own. There is a specially a significant improvement in the performance of all modalities for both classification tasks when compared to how CrisisNLP performed on its own before consolidation. For the informativeness task, the unimodal text-only model now is around 10% higher than the first run. The unimodal image-only model is now around 14% higher than the first run. The multimodal model sees the most improvement, being around 15% higher in most metrics while its F1 score is about 20% higher than the first run.

The humanitarian task sees higher improvement when compared to the informativeness task. The unimodal text-only model now is around 21% higher than the first run. The unimodal image-only model is now around 32% higher than the first run. The multimodal model is now around 29% higher than the first run.

4.4. Models' Performance for CrisisLex26

Tables 26 and 27 show the results for the informativeness and humanitarian category respectively for the CrisisLex26 with the different modalities in this study.

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|---------------|--------------|----------|-----------|--------|----------|
| Unimodal | Text | 76.2 | 75.9 | 76.0 | 76.0 |
| Unimodal | Image | 71.4 | 71.0 | 71.4 | 71.0 |
| Multimodal | Text + Image | 74.2 | 73.9 | 74.0 | 72.4 |

Table 26: Results for informativeness classification task for CrisisLex26

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|---------------|--------------|----------|-----------|--------|----------|
| Unimodal | Text | 55.2 | 48.1 | 55.0 | 48.2 |
| Unimodal | Image | 61.1 | 60.0 | 61.1 | 57.0 |
| Multimodal | Text + Image | 61.2 | 59.5 | 61.2 | 59.2 |

Table 27: Results for humanitarian category classification task for CrisisLex26

Similar to what was seen in the previous section with CrsisNLP, the results for the CrisisLex26 dataset are lower than CrisisMMD for both the informativeness and humanitarian category classification tasks. For the informativeness task, the unimodal text-only model here is around 5% less than CrisisMMD. The unimodal image-only model is around 12% less than CrisisMMD. The multimodal model here is around 10% lower as an average of most metrics, while the F1 score is around 12% less than CrisisMMD.

The performance of the model for the humanitarian category for all modalities is less than the results of CrisisMMD. The unimodal text-only model here is about 14% less than CrisisMMD for the accuracy and recall measures, while it is around 21% less for precision, and 18% less for F1 score. The unimodal image-only model is around 16% less than

CrisisMMD in most metrics, while the F1 score is around 20% less. The multimodal model is around 17% less than CrisisMMD.

The decrease in the performance for all modalities in CrisisLex26 is due to the much smaller dataset size as similarly seen in the previous section with CrisisNLP. 12708 records were available in CrisisMMD for the informativeness task, while only 975 records were available here in CrisisLex26. Similarly, 8079 records were available in CrisisMMD for the humanitarian category, while only 842 records were available here in CrisisLex26. A consistent finding here is that multimodal models perform better than the unimodal ones. This has been found to be true for all the three datasets for both classification tasks.

Further consolidation is applied to compensate the lower performance of CrisisLex26. The data from both CrisisMMD and CrisisNLP will be added to the data for CrisisLex26 to generate a much larger dataset. The models are trained using the consolidated dataset and tested using CrisisLex26. Then, the experiment is repeated and the models are rerun for both classification tasks. Tables 28 and 29 show the results for the informativeness and humanitarian category respectively for the CrisisLex26 consolidated with both CrisisMMD and CrisisNLP in with the different modalities.

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|---------------|--------------|----------|-----------|--------|----------|
| Unimodal | Text | 82.0 | 81.6 | 82.0 | 81.6 |
| Unimodal | Image | 83.7 | 83.5 | 83.7 | 83.6 |
| Multimodal | Text + Image | 84.7 | 84.5 | 84.2 | 84.5 |

Table 28: Results for informativeness classification task for Consolidated Dataset (CrisisMMD+CrisisNLP+CrisisLex26)

| Training mode | Modality | Accuracy | Precision | Recall | F1-score |
|---------------|--------------|----------|-----------|--------|----------|
| Unimodal | Text | 70.2 | 70.2 | 70.2 | 67.2 |
| Unimodal | Image | 77.5 | 77.1 | 77.5 | 77.0 |
| Multimodal | Text + Image | 77.9 | 77.3 | 77.8 | 77.3 |

Table 29: Results for humanitarian category task for Consolidated Dataset (CrisisMMD+CrisisNLP+CrisisLex26)

The results after the consolidation of all three datasets are better than results for the individual datasets on their own. There is a significant improvement in the performance of all modalities for both classification tasks when compared to how CrisisLex26 performed on its own before consolidation. For the informativeness task, the unimodal text-only model now is around 6% higher than the first run. The unimodal image-only model is now around 12% higher than the first run. The multimodal model is 10% higher in most metrics while its F1 score is about 12% higher than the first run.

The humanitarian task sees higher improvement when compared to the informativeness task. The unimodal text-only model now is around 15% higher than the first run for accuracy and recall measure, around 22% higher for precision, and around 19% higher for the F1 score. The unimodal image-only model is now around 16.5% higher than the first run in most metrics, while it is around 20% higher for the F1 score. The multimodal model is now around 17% higher than the first run.

The consolidation results in this section as well the previous one show that smaller datasets can be added to larger datasets of the same structures to improve the deep learning models' performance during classification. This is quite helpful since there is usually a lack of larger

datasets for disaster collections. This means that smaller disaster datasets including home-grown datasets in individual research papers can all be consolidated together after mapping their categories for consistency to lead to much improved performance results.

Chapter 5

Conclusions and Future Work

This chapter concludes this dissertation by providing an overview of the research undertaken, results and findings obtained, and future work in this research domain.

5.1. Conclusion

Social media is such an integral part of our everyday lives. People are using different social media platforms daily to constantly share and consume information. This plays a huge part during disaster situations, whether they are natural or man-made. In such times of crisis events, people need timely information to understand what is happening to stay safe. Social media platforms are the optimal channel for people to communicate during crisis events for their ease of access and rapid communications. The use of social media platforms during disasters help in quick broadcasting of crisis-related information to a much wider audience without having to wait for news agencies, facilitate tracking of affected individuals, simplify the process of asking for volunteers or donations, allow people to quickly ask for help, improve the effectiveness of rescue operations, and support information dissemination. Through analysis of disaster social media posts, this huge amount of disaster data available can be narrowed down to relevant categories to allow emergency services to manage their resources better for a more efficient disaster management response.

In this study, a multimodal deep learning system for automatic classification of disaster tweets was built. Two classification tasks were tackled which are informativeness (whether

the disaster tweet is informative or not informative) and the humanitarian category. A comparison between unimodal and multimodal deep learning models across three different representative disaster datasets (CrisisMMD, CrisisNLP, and CrisisLex26) was done. For text-only unimodal models, a CNN was used together with the word2vec word embeddings developed for Crisis informatics over 52 million disaster tweets (Imran, Mitra & Castillo 2016). For the image-only unimodal models, a VGG16 network is used with the last layer adapted to the classification tasks performed. Feature fusion is implemented for the multimodal model where two feature vectors from both the text and image modalities are obtained.

The first research question is whether it is possible to integrate multiple disaster datasets even if the labels are not identical in all of them, and how effective will the integration be if possible. The experiments done in this study show that such integration is possible and effective, leading to improved results compared to individual datasets. The humanitarian categories labelled in all three datasets are diverse with many variations and labels. In order to allow for uniform and comparable classification results, mapping between the different categories was done to have a consistent set of humanitarian categories that are used across all datasets.

Out of the three disaster datasets used, CrisisMMD has the largest size compared to CrisisNLP and CrisisLex26. Since deep learning models are data hungry, the performance of all models when applied to CrisisMMD were significantly better than with CrisisNLP or CrisisLex26. To compensate the lower performance of the smaller datasets, consolidation of datasets was done. First, CrisisNLP was merged with CrisisMMD, then all three datasets

were consolidated together. For the informativeness classification task, the unimodal text model in the consolidated dataset of CrisisMMD and CrisisNLP was better with an average of 10% than CrisisNLP on its own. Similarly, the unimodal image model improved by about 14%, and the multimodal model improved by around 15% in precision/accuracy/recall and around 20% in F1 score. A higher improvement was seen in the humanitarian category where the unimodal text model improved by around 21%, unimodal image model improved by around 32%, and the multimodal model improved by about 29%.

In the consolidated dataset for all three datasets (CrisisMMD + CrisisNLP + CrisisLex26), the performance of all models improved as well when compared to individual datasets. For the informativeness task, the best model got an F1 score of 84.5. The unimodal text model improved by 6% compared to how CrisisLex26 performed on its own before consolidation. The unimodal image model improved by around 12% and the multimodal improved by about 10% in all metrics except F1 score where it improved by 12%. For the humanitarian category task, the best model got an F1 score of 77.3. The unimodal text model in the consolidated dataset was 15% higher for recall and accuracy, 22% higher for precision and 19% higher for the F1 score. The unimodal image model improved by about 16.5% in all metrics except F1 score which improved by 20%. Furthermore, the multimodal model improved by around 17%.

These results further support the answer to the first research question, showing that mapping categories over multiple datasets for consistency and consolidating smaller datasets with larger ones significantly improves the deep learning models' performance during classification.

Since there is usually a lack of large labelled datasets in general and in the crisis informatics domain specifically, this mapping and consolidation approach can be utilized to make use of all the available smaller datasets for the deep learning models to achieve higher performance results. This consolidated dataset can now serve as a new baseline multimodal dataset after CrisisMMD.

Most of the studies demonstrated in the literature review were performed on small datasets that were mostly home-grown for specific disaster types. That has resulted in models that do not generalize well when they encounter new kinds of disasters. This issue was tackled here in this study with a larger consolidated dataset including a wide variety of natural and man-made disasters including typhoons, fires, earthquakes, tsunamis, etc.

All these findings support the answer to the second research question: How does performance of unimodal and multimodal models compare across different disaster datasets?

As the earlier results demonstrate, multimodal models perform better than unimodal models for all settings over informativeness and humanitarian category classification tasks across the three datasets, which is a finding supported by previous similar studies demonstrated in the literature review in chapter 2.

5.2. Limitations and Future Work

Social media data does not necessarily have a strong coupling between text and image content posted together. Sometimes, each of the modalities can be conveying a different information type that could contract the other one. So, it is essential to avoid the assumption that there is some strong relation between text and image content posted together on social

media. This aspect is not really explored much in the research community since all approaches for multimodal classification are built on the assumption that both the text and image modalities have a common label. Therefore, a possible future research challenge would be developing multimodal models that are based on text and image modalities with different labels.

This study tested mapping between categories and consolidation for three disaster datasets. A future research direction would be applying this mapping and consolidation approach to many more disaster datasets available in the crisis informatics research community. The target of developing the largest multimodal disaster dataset can then serve as a new baseline for further research directions in the field.

In addition, the mapping technique between humanitarian categories over different datasets could be further improved by a wider analysis of the most commonly represented labels over the numerous smaller datasets available, with a much more detailed semantic analysis of their representations, so that the distribution between categories is more balanced and all classes are equally represented.

Another possible future research direction is developing a disaster dashboard that analyzes social media data in real-time, categorizing the humanitarian categories and prioritizing them to allow emergency services to perform more efficient disaster management operations.

Variations of neural networks, embeddings and deep learning architectures could also be experimented with to examine possible roads to improvement. Examples could include variants of RNN models such as LSTM and BiLSTM for text classification, and different embeddings such as GloVe and ELMo contextual language embeddings to improve the

quality of dealing with informal text in social media data. Experiments with such model variants would allow investigations of which deep learning architectures perform best in the context of multimodal disaster data classification, and whether a generalized model can be developed to perform well on several disaster types.

References

- Abdelhade, N., Soliman, T. H. A. & Ibrahim, H. M. (2018). Detecting twitter users' opinions of arabic comments during various time episodes via deep neural network. *Advances in Intelligent Systems and Computing*, vol. 639, pp. 232–246.
- Acar, A. & Muraki, Y. (2011). Twitter for crisis communication: Lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*. Inderscience Publishers, vol. 7(3), pp. 392–402.
- Agarwal, M., Leekha, M., Sawhney, R. & Shah, R. R. (2020). Crisis-DIAS: Towards Multimodal Damage Analysis - Deployment, Challenges and Assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence (AAAI), vol. 34(01), pp. 346–353.
- Aggarwal, C. C. (2018). *Neural Networks and Deep Learning*. Neural Networks and Deep Learning. Springer International Publishing.
- Alam, F., Joty, S. & Imran, M. (2018). Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. *AAAI, (Icwsn)*, pp. 556–559.
- Alam, F., Ofli, F. & Imran, M. (2018). CrisisMMD: Multimodal twitter datasets from natural disasters. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, pp. 465–473.
- Alam, F., Ofli, F., Imran, M. & Aupetit, M. (2018). A twitter tale of three hurricanes: Harvey, Irma, and Maria. *Proceedings of the International ISCRAM Conference*. Information Systems for Crisis Response and Management, ISCRAM, pp. 553–572 [online]. [Accessed 14 April 2021]. Available at: <https://www.unocha.org/legacy/what-we-do/coordination-tools/cluster-coordination>.
- Alqhtani, S. M., Luo, S. & Regan, B. (2015). Fusing Text and Image for Event Detection in Twitter. *The International journal of Multimedia & Its Applications*. Academy and Industry Research Collaboration Center (AIRCC), vol. 7(1), pp. 27–35.
- Alshareef, H. N. & Grigoras, D. (2017). Using social media and the mobile cloud to enhance emergency and risk management. *Proceedings - 15th International Symposium on Parallel and Distributed Computing, ISPDC 2016*. Institute of Electrical and Electronics Engineers Inc., pp. 92–

99.

Ashktorab, Z., Brown, C., Nandi, M. & Culotta, A. (2014). Tweedr: Mining Twitter to Inform Disaster Response. *Proceedings of the 11th ISCRAM Conference*, pp. 354–358 [online]. [Accessed 18 April 2021]. Available at: <http://tweedr.dssg.io>.

Aswani, R., Kar, A. K., Ilavarasan, P. V. & Dwivedi, Y. K. (2018). Search engine marketing is not all gold: Insights from Twitter and SEOClerks. *International Journal of Information Management*. Elsevier Ltd, vol. 38(1), pp. 107–116.

Avvenuti, M., Cimino, M. G. C. A., Cresci, S., Marchetti, A. & Tesconi, M. (2016). A framework for detecting unfolding emergencies using humans as sensors. *SpringerPlus*. SpringerOpen, vol. 5(1), p. 43.

Baltrušaitis, T., Ahuja, C. & Morency, L.-P. (2018). ‘Challenges and applications in multimodal machine learning’. , in *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 2*. Association for Computing Machinery, pp. 17–48.

Baltrusaitis, T., Ahuja, C. & Morency, L. P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE Computer Society, pp. 423–443.

Basnyat, B., Anam, A., Singh, N., Gangopadhyay, A. & Roy, N. (2017). Analyzing Social Media Texts and Images to Assess the Impact of Flash Floods in Cities. *2017 IEEE International Conference on Smart Computing, SMARTCOMP 2017*. Institute of Electrical and Electronics Engineers Inc.

Blank, G. & Reisdorf, B. C. (2012). THE PARTICIPATORY WEB. *Information, Communication & Society*. Nova Science Publishers, Inc., vol. 15(4), pp. 537–554.

Boccignone, G., Bortoletto, R., Zhang, Y., Lagerstrom, R., Arzhaeva, Y., Szul, P., Obst, O., Power, R., Robinson, B. & Bednarz, T. (2016). Image classification to Support emergency Situation Awareness. *Frontiers in Robotics and AI*, p. 54.

Bodenhamer, M. (2011). *HYOGO FRAMEWORK FOR ACTION 2005-2015 Building the Resilience of Nations and Communities to Disasters*.

Brownlee, J. (2020). *Why do I Get Different Results Each Time in Machine Learning? Machine Learning Mastery* [online]. Available at: <https://machinelearningmastery.com/different-results-each-time-in-machine-learning/>.

Budak, C., Agrawal, D. & El Abbadi, A. (2012). Diffusion of information in social networks: Is it all local? *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 121–130.

Burel, G. & Alani, H. (2018). Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media.

Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H.-W., Mitra, P., Wu, D., Tapia, A. H., Giles, L., Jansen, B. J. & Yen, J. (2011). *Classifying Text Messages for the Haiti Earthquake* [online]. [Accessed 14 April 2021]. Available at: <http://haiti.ushahidi.com>.

Caragea, C., Silvescu, A. & Tapia, A. H. (2016). Identifying Informative Messages in Disaster Events using Convolutional Neural Networks. *International Conference on Information Systems for Crisis Response and Management*, pp. 137–147 [online]. [Accessed 18 April 2021]. Available at: <http://t.co/XW4Fn27tVy>.

Cassa, C. A., Chunara, R., Mandl, K. & Brownstein, J. S. (2013). Twitter as a Sentinel in Emergency Situations: Lessons from the Boston Marathon Explosions. *PLoS Currents*. Public Library of Science, vol. 5(JUNE).

Chaudhuri, N. & Bose, I. (2019). Application of Image Analytics for Disaster Response in Smart Cities. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences.

Chen, T., Lu, D., Kan, M. Y. & Cui, P. (2013). Understanding and classifying image tweets. *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*. New York, New York, USA: ACM Press, pp. 781–784.

CisisLex: Crisis Collections. (n.d.) [online]. Available at: <http://crisislex.org/data-collections.html#CrisisLexT26>.

CRED. (2020). *Human Cost of Disasters (2000-2019) - Cred Crunch from "EM-DAT: The OFDA/CRED International Disaster Database"*.

CRED. (n.d.). *EM-DAT: The International Disaster Database*. Centre for Research on the

Epidemiology of Disasters (CRED) [online]. Available at: <https://www.emdat.be/>.

CrisisLex GitHub Repository. (n.d.) [online]. Available at:
<https://github.com/sajao/CrisisLex/tree/master/data/CrisisLexT26/>.

Dean, B. (2021). *How Many People Use Twitter in 2021? [New Twitter Stats]* [online]. [Accessed 12 April 2021]. Available at: <https://backlinko.com/twitter-users>.

Deller, R. (2011). Twittering on: Audience research and participation using Twitter. *Participations: Journal of Audience & Reception Studies*, vol. 8(1), pp. 216–245.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li & Li Fei-Fei. (2010). ImageNet: A large-scale hierarchical image database. Institute of Electrical and Electronics Engineers (IEEE), pp. 248–255.

Dewan, P., Suri, A., Bharadhwaj, V., Mithal, A. & Kumaraguru, P. (2017). Towards understanding crisis events on online social networks through pictures. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*. New York, NY, USA: Association for Computing Machinery, Inc, pp. 439–446.

Dong, H., Halem, M. & Zhou, S. (2013). Social media data analytics applied to Hurricane Sandy. *Proceedings - SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013*, pp. 963–966.

Fabrega, J. & Paredes, P. (2013). Social Contagion and Cascade Behaviors on Twitter. *Information*. MDPI AG, vol. 4(2), pp. 171–181.

Formentin, M., Bortree, D. and Fraustino, J. D. (2012). Navigating anger in happy valley: using facebook for crisis response and image repair in the wake of the Sandusky scandal. *Meeting of the Association for Education in Journalism and Mass Communication, Chicago, IL*.

Fraustino, J. D., Brooke, L. & Yan, J. (2012). Social Media Use during Disasters: A Review of the Knowledge Base and Gaps. *Final Report to Human Factors/Behavioral Sciences Division, Science and Technology Directorate, U.S. Department of Homeland Security. College Park, MD: START* [online]. Available at:
https://www.start.umd.edu/sites/default/files/files/publications/START_SocialMediaUseduringDisasters_LitReview.pdf.

Fujii, Y., Satake, K., Sakai, S., Shinohara, M. & Kanazawa, T. (2011). Tsunami source of the 2011 off the Pacific coast of Tohoku Earthquake. *Earth, Planets and Space*. Springer Berlin, vol. 63(7),

pp. 815–820.

Fung, I. C. H., Duke, C. H., Finch, K. C., Snook, K. R., Tseng, P. L., Hernandez, A. C., Gambhir, M., Fu, K. W. & Tse, Z. T. H. (2016). Ebola virus disease and social media: A systematic review. *American Journal of Infection Control*. Mosby Inc., pp. 1660–1671.

Gautam, A. K., Misra, L., Kumar, A., Misra, K., Aggarwal, S. & Shah, R. R. (2019). Multimodal analysis of disaster tweets. *Proceedings - 2019 IEEE 5th International Conference on Multimedia Big Data, BigMM 2019*. IEEE, pp. 94–103.

gensim. (n.d.) [online]. Available at: <https://pypi.org/project/gensim/>.

Girtelschmid, S., Salfinger, A., Proll, B., Retschitzegger, W. & Schwinger, W. (2016). Near real-time detection of crisis situations. *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2016 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp. 247–252.

Goldberg, Y. (2015). A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*. AI Access Foundation, vol. 57, pp. 345–420 [online]. [Accessed 24 May 2021]. Available at: <http://arxiv.org/abs/1510.00726>.

Goutte, C. & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *Lecture Notes in Computer Science*. Springer Verlag, pp. 345–359.

Gupta, A., Lamba, H., Kumaraguru, P. & Joshi, A. (2013). Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*. New York, New York, USA: ACM Press [online]. [Accessed 26 May 2021]. Available at: <http://www.guardian.co.uk/world/us-news->

Hagar, C. (2010). ‘Crisis informatics’. , in *Encyclopedia of Information Science and Technology* Third Edit. IGI Global, pp. 1350–1358.

Hagen, L., Keller, T., Neely, S., DePaula, N. & Robert-Cooperman, C. (2018). Crisis Communications in the Age of Social Media. *Social Science Computer Review*. SAGE Publications Inc., vol. 36(5), pp. 523–541.

Honan, M. (2011). *Watch The Virginia Earthquake Spread Across Twitter*. *Gizmodo* [online].

[Accessed 7 April 2021]. Available at: <https://gizmodo.com/watch-the-virginia-earthquake-spread-across-twitter-5834048>.

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

Hussain, M., Bird, J. J. & Faria, D. R. (2019). A study on CNN transfer learning for image classification. *Advances in Intelligent Systems and Computing*. Springer Verlag, pp. 191–202.

Ilyas, A. (2014). MicroFilters: Harnessing twitter for disaster management. *Proceedings of the 4th IEEE Global Humanitarian Technology Conference, GHTC 2014*. Institute of Electrical and Electronics Engineers Inc., pp. 417–424.

Imran, M. & Castillo, C. (2015). Towards a data-driven approach to identify crisis-related topics in social media streams. *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*. New York, NY, USA: Association for Computing Machinery, Inc, pp. 1205–1210.

Imran, M., Castillo, C., Diaz, F. & Vieweg, S. (2015). Processing social media messages in Mass Emergency: A survey. *ACM Computing Surveys*. Association for Computing Machinery, vol. 47(4).

Imran, M., Elbassuoni, S., Castillo QCRI, C., Diaz, F. & Meier QCRI, P. (2013). *Extracting Information Nuggets from Disaster-Related Messages in Social Media* [online]. [Accessed 6 April 2021]. Available at: <http://www.crowdfunder.com>.

Imran, M., Mitra, P. & Castillo, C. (2016). Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pp. 1638–1643.

Imran, M., Ofli, F., Caragea, D. & Torralba, A. (2020). Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. *Information Processing and Management*, vol. 57(5), pp. 1–9.

Ioffe, S. & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167*.

- Jeni, L. A., Cohn, J. F. & De La Torre, F. (2013). Facing imbalanced data - Recommendations for the use of performance metrics. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pp. 245–251.
- Jomaa, H. S., Rizk, Y. & Awad, M. (2017). Semantic and Visual Cues for Humanitarian Computing of Natural Disaster Damage Images. *Proceedings - 12th International Conference on Signal Image Technology and Internet-Based Systems, SITIS 2016*. Institute of Electrical and Electronics Engineers Inc., pp. 404–411.
- Kalliatakis, G., Ehsan, S., Fasli, M. & Mcdonald-Maier, K. (2019). DisplaceNet: Recognising Displaced People from Images by Exploiting Dominance Level. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 33–38.
- Kaur, T. & Gandhi, T. K. (2020). Deep convolutional neural networks with transfer learning for automated brain image classification. *Machine Vision and Applications*. Springer, pp. 1–16.
- Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., Natsev, A. & Xie, L. (2012). Social media use by government: From the routine to the critical. *Government Information Quarterly*. JAI, vol. 29(4), pp. 480–491.
- Keim, M. E. & Noji, E. (2011). Emergent use of social media: a new age of opportunity for disaster resilience. *American journal of disaster medicine*, vol. 6(1), pp. 47–54.
- Kemp, S. (2021). *Digital 2021: Global Overview Report — DataReportal – Global Digital Insights* [online]. [Accessed 7 April 2021]. Available at: <https://datareportal.com/reports/digital-2021-global-overview-report>.
- Keras. (n.d.) [online]. Available at: <https://keras.io/>.
- Krishna, S. T. & Kumar Kalluri, H. (2019). Deep Learning and Transfer Learning Approaches for Image Classification. *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 7(5S4), pp. 427–432 [online]. [Accessed 23 May 2021]. Available at: <https://www.researchgate.net/publication/333666150>.
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems*, pp. 1097–1105 [online]. [Accessed 23 May 2021]. Available at: <http://code.google.com/p/cuda-convnet/>.

- Kumar, A., Singh, J. P., Dwivedi, Y. K. & Rana, N. P. (2020). A deep multi-modal neural network for informative Twitter content classification during emergencies. *Annals of Operations Research*. Springer, pp. 1–32.
- Kumar, S., Morstatter, F., Zafarani, R. & Liu, H. (2013). Whom should I follow? Identifying relevant users during crises. *HT 2013 - Proceedings of the 24th ACM Conference on Hypertext and Social Media*. New York, New York, USA: ACM Press, pp. 139–147.
- Kushwaha, V., Singh, M., Singh, R., Vatsa, M., Ratha, N. & Chellappa, R. (2018). Disguised Faces in the Wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1–9.
- Kwon, J. & Han, I. (2013). Information diffusion with content crossover in online social media: An empirical analysis of the social transmission process in Twitter. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 3292–3301.
- Lai, C. H., She, B. & Tao, C. C. (2017). Connecting the dots: A longitudinal observation of relief organizations' representational networks on social media. *Computers in Human Behavior*. Elsevier Ltd, vol. 74, pp. 224–234.
- Landwehr, P. M., Wei, W., Kowalchuck, M. & Carley, K. M. (2016). Using tweets to support disaster planning, warning and response. *Safety Science*. Elsevier B.V., vol. 90, pp. 33–47.
- Laylavi, F., Rajabifard, A. & Kalantari, M. (2017). Event relatedness assessment of Twitter messages for emergency response. *Information Processing and Management*. Elsevier Ltd, vol. 53(1), pp. 266–280.
- Li, Z. & Di, S. (2017). Preparation and properties of micro-arc oxidation self-lubricating composite coatings containing paraffin. *Journal of Alloys and Compounds*. Elsevier Ltd, vol. 719, pp. 1–14.
- Lobb, A., Mock, N. & Hutchinson, P. L. (2012). Traditional and social media coverage and charitable giving following the 2010 earthquake in Haiti. *Prehospital and Disaster Medicine*, vol. 27(4), pp. 319–324.
- Martínez-Rojas, M., Pardo-Ferreira, M. del C. & Rubio-Romero, J. C. (2018a). Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management*. Elsevier, vol. 43(April), pp. 196–208.

- Martínez-Rojas, M., Pardo-Ferreira, M. del C. & Rubio-Romero, J. C. (2018b). Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management*. Elsevier Ltd, pp. 196–208.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR [online]. [Accessed 23 May 2021]. Available at: <http://ronan.collobert.com/senna/>.
- Mills, A., Chen, R., Lee, J. & Raghav Rao, H. (2009). Web 2.0 Emergency Applications: How Useful Can Twitter be for Emergency Response? *Journal of Information Privacy and Security*. Informa UK Limited, vol. 5(3), pp. 3–26.
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. & Gao, J. (2021). Deep Learning--based Text Classification. *ACM Computing Surveys*. ACM PUB27 New York, NY, USA , vol. 54(3), pp. 1–40.
- Mirbabaie, M., Bunker, D., Stieglitz, S., Marx, J. & Ehnis, C. (2020). Social media in times of crisis: Learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response. *Journal of Information Technology*. SAGE Publications Ltd, vol. 35(3), pp. 195–213.
- Mouzannar, H., Rizk, Y. & Awad, M. (2018). *Damage Identification in Social Media Posts using Multimodal Deep Learning*.
- Muralidharan, S., Dillistone, K. & Shin, J. H. (2011). The Gulf Coast oil spill: Extending the theory of image restoration discourse to the realm of social media and beyond petroleum. *Public Relations Review*. JAI, vol. 37(3), pp. 226–232.
- Natural Language ToolKit (nltk)*. (n.d.) [online]. Available at: <https://www.nltk.org/>.
- Nazer, T. H., Xue, G., Ji, Y. & Liu, H. (2017). Intelligent disaster response via social media analysis - A survey. *arXiv*, vol. 19(1), pp. 46–59.
- Nguyen, D. T., Alam, F., Ofli, F. & Imran, M. (2017). Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises. *Proceedings of the International ISCRAM Conference*. Information Systems for Crisis Response and Management, ISCRAM, vol. 2017-May, pp. 499–511 [online]. [Accessed 14 April 2021]. Available at: <http://arxiv.org/abs/1704.02602>.

- Nguyen, D. T., Joty, S., Imran, M., Sajjad, H. & Mitra, P. (2016). Applications of Online Deep Learning for Crisis Response Using Social Media Information [online]. [Accessed 18 April 2021]. Available at: <http://arxiv.org/abs/1610.01030>.
- Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M. & Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, (Icws), pp. 632–635.
- Nguyen, D. T., Ofli, F., Imran, M. & Mitra, P. (2017). Damage assessment from social media imagery data during disasters. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*. Association for Computing Machinery, Inc, pp. 569–576.
- Noreña, D., Akhavan-Tabatabaei, R., Yamín, L. & Ospina, W. (2011). Using discrete event simulation to evaluate the logistics of medical attention during the relief operations in an earthquake in Bogota. *Proceedings - Winter Simulation Conference*, pp. 2661–2673.
- NumPy. (n.d.) [online]. Available at: <https://numpy.org/>.
- Ofli, F., Alam, F. & Imran, M. (2020). Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. *arXiv*, vol. 1(May 2020).
- Olteanu, A., Vieweg, S. & Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 994–1009.
- Palen, L. & Anderson, K. M. (2016). Crisis informatics—New data for extraordinary times. *Science*, vol. 353(6296).
- Palen, L., Vieweg, S., Liu, S. B. & Hughes, A. L. (2009). Crisis in a Networked World. *Social Science Computer Review*. SAGE PublicationsSage CA: Los Angeles, CA, vol. 27(4), pp. 467–480.
- Panagiotopoulos, P., Barnett, J., Bigdeli, A. Z. & Sams, S. (2016). Social media in emergency management: Twitter as a tool for communicating risks to the public. *Technological Forecasting and Social Change*. Elsevier Inc., vol. 111, pp. 86–96.

Pandas. (n.d.) [online]. Available at: <https://pandas.pydata.org/>.

Perng, S.-Y., Büscher, M., Wood, L., Halvorsrud, R., Stiso, M., Ramirez, L. & Al-Akkad, A. (2013). Peripheral Response: Microblogging During the 22/7/2011 Norway Attacks. *International Journal of Information Systems for Crisis Response and Management*, vol. 5(1), pp. 41–57.

Piedra, N., Chicaiza, J. & Torres-Guarnizo, D. (2017). Caracterización de eventos naturales y epidemias registradas en Twitter: Fenómeno del Niño, Zika y Chikungunya. *Iberian Conference on Information Systems and Technologies, CISTI*. IEEE Computer Society.

Pogrebnyakov, N. & Maldonado, E. (2018). Didn't roger that: Social media message complexity and situational awareness of emergency responders. *International Journal of Information Management*. Elsevier Ltd, vol. 40, pp. 166–174.

Purohit, H., Castillo, C., Imran, M. & Pandev, R. (2018). Social-EOC: Serviceability model to rank social media requests for emergency operation centers. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*. Institute of Electrical and Electronics Engineers Inc., pp. 119–126.

Radianti, J., Hiltz, S. R. & Labaka, L. (2016). An overview of public concerns during the recovery period after a major earthquake: Nepal twitter analysis. *Proceedings of the Annual Hawaii International Conference on System Sciences*. IEEE Computer Society, pp. 136–145.

Ragini, J. R., Anand, P. M. R. & Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*. Elsevier Ltd, vol. 42, pp. 13–24.

Ray, A. & Bala, P. K. (2020). Social media for improved process management in organizations during disasters. *Knowledge and Process Management*, vol. 27(1), pp. 63–74.

Reuter, C., Hughes, A. L. & Kaufhold, M.-A. (2018). Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *International Journal of Human-Computer Interaction*. Taylor and Francis Inc., vol. 34(4), pp. 280–294.

Reuter, C. & Kaufhold, M.-A. (2018). Fifteen years of social media in emergencies: A retrospective review and future directions for crisis Informatics. *Journal of Contingencies and Crisis Management*. Blackwell Publishing Ltd, vol. 26(1), pp. 41–57.

- Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). ‘Why should i trust you?’ Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, pp. 1135–1144.
- Rizk, Y., Awad, M., Jomaa, H. S. & Castillo, C. (2019). A computationally efficient multi-modal classification approach of disaster-related Twitter images. *Proceedings of the ACM Symposium on Applied Computing*. New York, NY, USA: Association for Computing Machinery, pp. 2050–2059.
- Romero, D. M., Meeder, B. & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*. New York, New York, USA: ACM Press, pp. 695–704.
- El Sayed, M. J. (2020). Beirut ammonium nitrate explosion: A man-made disaster in times of CoVID19 pandemic. *Disaster Medicine and Public Health Preparedness*. Cambridge University Press, pp. 1–5.
- scikit-learn*. (n.d.) [online]. Available at: <https://scikit-learn.org/stable/>.
- Shaluf, I. M. (2007). Disaster types. *Disaster Prevention and Management: An International Journal*. Emerald Group Publishing Ltd., vol. 16(5), pp. 704–717.
- Simon, T., Goldberg, A. & Adini, B. (2015). Socializing in emergencies - A review of the use of social media in emergency situations. *International Journal of Information Management*. Elsevier Ltd, pp. 609–619.
- Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR [online]. [Accessed 23 May 2021]. Available at: <http://www.robots.ox.ac.uk/>.
- Soares, M. de O., Teixeira, C. E. P., Bezerra, L. E. A., Paiva, S. V., Tavares, T. C. L., Garcia, T. M., de Araújo, J. T., Campos, C. C., Ferreira, S. M. C., Matthews-Cascon, H., Frota, A., Mont’Alverne, T. C. F., Silva, S. T., Rabelo, E. F., Barroso, C. X., Freitas, J. E. P. de, Melo Júnior, M. de, Campelo, R. P. de S., Santana, C. S. de, Carneiro, P. B. de M., Meirelles, A. J., Santos, B. A., Oliveira, A. H. B. de, Horta, P. & Cavalcante, R. M. (2020). Oil spill in South Atlantic (Brazil):

Environmental and governmental disaster. *Marine Policy*. Elsevier Ltd, vol. 115, p. 103879.

Sokolova, M., Japkowicz, N. & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *AAAI Workshop - Technical Report*. Springer, Berlin, Heidelberg, pp. 24–29.

Spiro, E. S., Sutton, J., Greczek, M., Fitzhugh, S., Pierski, N. & Butts, C. T. (2012). Rumoring during extreme events: A case study of Deepwater Horizon 2010. *Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12*. New York, New York, USA: Association for Computing Machinery, pp. 275–283.

Squicciarini, A., Tapia, A. & Stehle, S. (2017). Sentiment analysis during Hurricane Sandy in emergency response. *International Journal of Disaster Risk Reduction*. Elsevier Ltd, vol. 21, pp. 213–222.

Srivastava, N., Hinton, G., Krizhevsky, A. & Salakhutdinov, R. (2014). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. *Journal of Machine Learning Research*.

Starbird, K., Palen, L., Hughes, A. L. & Vieweg, S. (2010). Chatter on The Red: What hazards threat reveals about the social life of microblogged information. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. New York, New York, USA: ACM Press, pp. 241–250.

Sutton, J., Palen, L. & Shklovski, I. (2008). *Backchannels on the Front Lines: Emergent Uses of Social Media in the 2007 Southern California Wildfires*.

Takahashi, B., Tandoc, E. C. & Carmichael, C. (2015). Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines. *Computers in Human Behavior*. Elsevier Ltd, vol. 50, pp. 392–398.

Tan, M. L., Prasanna, R., Stock, K., Hudson-Doyle, E., Leonard, G. & Johnston, D. (2017). Mobile applications in crisis informatics literature: A systematic review. *International Journal of Disaster Risk Reduction*. Elsevier Ltd, vol. 24, pp. 297–311.

Tatsubori, M., Watanabe, H., Shibayama, A., Sato, S. & Imamura, F. (2012). Social Web in disaster archives. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion*. New York, New York, USA: ACM Press, pp. 715–716.

- Taylor, M., Wells, G., Howell, G. and Raphael, B. (2012). The role of social media as psychological first aid as a support to community resilience building. *The Australian Journal of Emergency Management*, vol. 27(1), pp. 20–26.
- Thom, D., Kruger, R., Ertl, T., Bechstedt, U., Platz, A., Zisgen, J. & Volland, B. (2015). Can twitter really save your life? A case study of visual social media analytics for situation awareness. *IEEE Pacific Visualization Symposium*. IEEE Computer Society, pp. 183–190.
- Todd, D. & Todd, H. (2011). *Natural Disaster Response Lessons from Evaluations of the World Bank and Others*. World Bank, Washington, DC [online]. [Accessed 6 April 2021]. Available at: <http://ieg.worldbankgroup.org>.
- Twitter, Inc. (2021) [online]. [Accessed 12 April 2021]. Available at: <https://investor.twitterinc.com/home/default.aspx>.
- Vieweg, S., Castillo, C. & Imran, M. (2014). Integrating social media communications into the rapid assessment of sudden onset disasters. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 444–461.
- Vitale, D., Ferragina, P. & Scaiella, U. (2012). Classification of short texts by deploying topical annotations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Berlin, Heidelberg, pp. 376–387.
- Wang, Z., Ye, X. & Tsou, M. H. (2016). Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*. Springer Netherlands, vol. 83(1), pp. 523–540.
- Williams, R., Williams, G. & Burton, D. (2012). *The Use of Social Media for Disaster Recovery*.
- Wu, H., Liu, Y. & Wang, J. (2020). Review of text classification methods on deep learning. *Computers, Materials and Continua*, vol. 63(3), pp. 1309–1321.
- Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*. Neural information processing systems foundation, vol. 4(January), pp. 3320–3328 [online]. [Accessed 23 May 2021]. Available at: <http://arxiv.org/abs/1411.1792>.
- Yu, M., Huang, Q., Qin, H., Scheele, C. & Yang, C. (2019). Deep learning for real-time social

media text classification for situation awareness – using Hurricanes Sandy, Harvey, and Irma as case studies. *International Journal of Digital Earth*. Taylor and Francis Ltd., vol. 12(11), pp. 1230–1247.

Yu, M., Li, Z., Yu, Z., He, J. & Zhou, J. (2020). Communication related health crisis on social media: a case of COVID-19 outbreak. *Current Issues in Tourism*. Routledge, pp. 1–7.

Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M. & Starbird, K. (2018). From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *Proceedings of the ACM on Human-Computer Interaction*. Association for Computing Machinery, vol. 2(CSCW), pp. 1–18.

Zeiler, M. D. (2012). ADADELTA: An Adaptive Learning Rate Method [online]. [Accessed 23 May 2021]. Available at: <http://arxiv.org/abs/1212.5701>.

Zhou, Q., Huang, W. & Zhang, Y. (2011). Identifying critical success factors in emergency management using a fuzzy DEMATEL method. *Safety Science*, vol. 49(2), pp. 243–252.

Zhou, Y. (2020). A Review of Text Classification Based on Deep Learning. *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp. 132–136.

Zoppi, T., Ceccarelli, A., Lollini, P., Bondavalli, A., Lo Piccolo, F., Giunta, G. & Morreale, V. (2016). Presenting the Proper Data to the Crisis Management Operator: A Relevance Labelling Strategy. *Proceedings of IEEE International Symposium on High Assurance Systems Engineering*. IEEE Computer Society, pp. 228–235.

Appendix A

For the informativeness classification task over CrisisMMD, the performance of all models for all the three modalities done in this study is consistent with the results found by (Ofli, Alam & Imran 2020). The highest difference of 0.6% is between the accuracy for the unimodal text model in this study and (Ofli, Alam & Imran 2020). The unimodal image model in this study is 0.3% better, while the multimodal model in the study is only 0.1% higher.

For the humanitarian category classification task, the performance of all models for all the three modalities done in this study is also consistent with the results found by (Ofli, Alam & Imran 2020). The unimodal text model in the authors' experiment had better results of about 1%. The unimodal image in this study is 0.6% higher than the ones in (Ofli, Alam & Imran 2020). For the multimodal model, the accuracy in their experiment is 0.9% higher, precision is 1.7% higher, recall is 0.5 higher, and the F1 score is 1.4% higher than in this study. The multimodal model has higher differences between the two experiments compared to the other two unimodal models, but the differences are minimal and they are still consistent with each other.

The minimal differences between the two experiments can be due to the difference in some of the data instances. When this study was performed and the tweets were retrieved, some tweets were not available because of suspended user accounts or broken image links. Those could have still been available when the experiment in (Ofli, Alam & Imran 2020) was done. Another reason is the differences caused by the environment the models are run on in both

studies. In (Ofli, Alam & Imran 2020), they were using Python 2.7 but this study uses the much newer version Python 3.7.10. Many of the Python machine learning libraries used in this study are also the updated newer versions when compared to the other experiment as shown in Table 30.

| Python Library | (Ofli, Alam & Imran 2020) | This Study |
|---|--------------------------------------|-------------------|
| Keras (<i>Keras</i> n.d.) | 2.2.4 | 2.5.0 |
| pandas (<i>Pandas</i> n.d.) | 0.24.2 | 1.1.5 |
| nltk (<i>Natural Language ToolKit (nltk)</i> n.d.) | 3.4 | 3.5 |
| gensim (<i>gensim</i> n.d.) | 3.7.3 | 3.8.3 |
| numpy (<i>NumPy</i> n.d.) | 1.14.2 | 1.19.5 |
| scikit-learn (<i>scikit-learn</i> n.d.) | 0.22.1 | 0.22.2 |

Table 30: Machine Learning Python Libraries in both studies

Differences in libraries' versions can sometimes change the functionality, therefore that could play a factor in different results for the same model (Brownlee 2020).