

Applications of deep learning in extracting actionable information from
crisis-related social media content

by

Sarthak Khanal

B.E., Tribhuvan University, Nepal, 2015

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Abstract

We have witnessed a large number of crisis situations in recent years, from natural disasters to man-made disasters and also to deadly animal and human health crises, culminating with the ongoing Covid-19 public health crisis. Disasters can have devastating health and socio-economic impacts. Emergency response and critical resource management during crises are pivotal tasks in mitigating the impacts of such events. These tasks require time-critical and reliable information for effective implementation. During emergent crises, there is a huge influx of information from various sources, which makes the task of collecting and managing reliable information harder. Identifying key information relevant for emergency responders and policy makers from huge streams of data is an infeasible task for human to attempt. There is a clear need of a pipeline of systems that can monitor, identify and collect actionable and relevant information from incomplete and noisy sources of data.

Social media has evolved into a platform for people to share their concerns, report information as eyewitnesses of events, and also call for help, especially during crisis situations. However, due to the unstructured nature of data shared in these digital media, inherent noise and potential misinformation, extraction of actionable information is a challenging task. Considering the challenges associated with modern data-driven emergency response and crisis management, deep-learning is a natural choice in making use of the large volume of unstructured data. However, deep-learning models, typically, require a large amount of annotated or labelled data, which may not always be available for an emergent crisis. This dissertation aims to address some of these issues by exploring multi-task and multimodal deep-learning approaches, combined with self-supervised representation learning.

From an application point of view, this dissertation tackles two specific tasks surrounding crisis information management: firstly, the time-critical task of identifying actionable information for emergent crisis, and secondly, the task of analyzing public response to crisis events and the policies surrounding the events through social-media.

Applications of deep learning in extracting actionable information from
crisis-related social media content

by

Sarthak Khanal

B.E., Tribhuvan University, Nepal, 2015

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Approved by:

Major Professor
Doina Caragea

Copyright

© Sarthak Khanal 2022.

Abstract

We have witnessed a large number of crisis situations in recent years, from natural disasters to man-made disasters and also to deadly animal and human health crises, culminating with the ongoing Covid-19 public health crisis. Disasters can have devastating health and socio-economic impacts. Emergency response and critical resource management during crises are pivotal tasks in mitigating the impacts of such events. These tasks require time-critical and reliable information for effective implementation. During emergent crises, there is a huge influx of information from various sources, which makes the task of collecting and managing reliable information harder. Identifying key information relevant for emergency responders and policy makers from huge streams of data is an infeasible task for human to attempt. There is a clear need of a pipeline of systems that can monitor, identify and collect actionable and relevant information from incomplete and noisy sources of data.

Social media has evolved into a platform for people to share their concerns, report information as eyewitnesses of events, and also call for help, especially during crisis situations. However, due to the unstructured nature of data shared in these digital media, inherent noise and potential misinformation, extraction of actionable information is a challenging task. Considering the challenges associated with modern data-driven emergency response and crisis management, deep-learning is a natural choice in making use of the large volume of unstructured data. However, deep-learning models, typically, require a large amount of annotated or labelled data, which may not always be available for an emergent crisis. This dissertation aims to address some of these issues by exploring multi-task and multimodal deep-learning approaches, combined with self-supervised representation learning.

From an application point of view, this dissertation tackles two specific tasks surrounding crisis information management: firstly, the time-critical task of identifying actionable information for emergent crisis, and secondly, the task of analyzing public response to crisis events and the policies surrounding the events through social-media.

Table of Contents

List of Figures	x
List of Tables	xii
Acknowledgements	xv
Dedication	xvi
1 Introduction	1
1.1 Social media as a data source	2
1.2 Deep learning	2
1.3 Contributions of this dissertation	4
1.3.1 Data collection	4
1.3.2 Identification of fine-grained location mentions in crisis tweets	5
1.3.3 Multi-task learning to enable location mention identification in the early hours of a crisis event	7
1.3.4 Contrastive learning for multimodal classification of crisis related tweets . .	8
1.3.5 Using content analysis and machine learning to identify COVID-19 infor- mation relevant to low-income households on social media	9
1.3.6 Summary	9
2 Identification of fine-grained location mentions in crisis tweets	11
2.1 Introduction	11
2.2 Related work	14
2.2.1 NER and location mention identification	14

2.2.2	Fine-grained location types	15
2.2.3	State-of-the-art approaches for NER	16
2.2.4	Location datasets	16
2.3	Datasets	17
2.3.1	Data collection	17
2.3.2	Data annotation and quality assessment	17
2.3.3	Benchmark datasets	19
2.4	Background and approaches	20
2.4.1	BERT embeddings	21
2.4.2	BiLSTM	21
2.4.3	Conditional random fields (CRFs)	22
2.4.4	Multi-task learning	22
2.4.5	Proposed model and variants	23
2.5	Experimental setup	24
2.5.1	Baseline	24
2.5.2	Metrics	24
2.5.3	Implementation details	24
2.5.4	Experiments	25
2.6	Results and discussion	26
2.6.1	In-domain setting	26
2.6.2	Cross-domain setting	27
2.6.3	Error analysis	28
2.7	Conclusions	29
3	Multi-task learning to enable location mention identification in the early hours of a crisis event	30
3.1	Introduction	30
3.2	Related work	32

3.3	Background and approaches	33
3.3.1	Multi-task learning	33
3.3.2	Token classification tasks	33
3.3.3	Sequence classification tasks	34
3.3.4	Shared module	34
3.4	Experimental setup	35
3.4.1	Datasets	35
3.4.2	Baseline and metrics	36
3.4.3	Experiments and implementation details	36
3.5	Results and discussion	37
3.6	Conclusions	39
4	Contrastive learning for multimodal classification of crisis related tweets	40
4.1	Introduction	40
4.2	Related work	42
4.3	Background and problem definition	44
4.4	Datasets	45
4.4.1	CrisisMMD	45
4.4.2	Pre-training dataset	46
4.5	Experiments	47
4.5.1	Baseline and proposed models	47
4.5.2	Performance metrics	48
4.6	Results and discussions	48
4.6.1	CLIP variants (pre-trained models)	48
4.6.2	CrisisMMD tasks	49
4.7	Conclusions	52

5	Using content analysis and machine learning to identify COVID-19 information relevant to low-income households on social media	53
5.1	Introduction	53
5.2	Theoretical frameworks for content analysis	56
5.2.1	Crisis communication framework	56
5.2.2	Social order theory	57
5.2.3	Research questions	57
5.3	Procedure of content analysis	58
5.3.1	Coding categories and coding protocol	58
5.3.2	Sample data	63
5.3.3	Coding process and interoperability	63
5.3.4	Annotated dataset and insights provided	64
5.4	Automated ML models for identifying LIH information on social media	66
5.4.1	Benchmark dataset for ML	66
5.4.2	Model design	67
5.4.3	Metrics	68
5.4.4	Experimental setup	69
5.5	ML results and discussion	71
5.6	Conclusions	72
6	Concluding remarks and future prospects	74
	Bibliography	77

List of Figures

1.1	Proposed Twitter data streaming pipeline.	5
1.2	Overview of the overarching goal for the identification of fine-grained location mentions in crisis tweets. The scope of the project is highlighted in yellow oval. . .	6
1.3	Multi-task learning architecture containing a shared transformer module (γ) and k task-specific modules (τ). $X_i(1 \leq i \leq k)$ are the datasets used for each task, respectively.	7
1.4	Overview of the two-step multimodal training process with contrastive pre-training and task-specific fine-tuning.	8
2.1	Overview of the the deep learning architecture for fine-grained location identification.	14
2.2	Entity distribution by number of tokens in a location in the <i>Mixed</i> and <i>Covid</i> datasets, respectively.	19
2.3	Proposed model architecture. The model can be seen as an MTL model, with two objective corresponding to two tasks. The primary task (right) is to predict fine-grained location tags, while the auxiliary task (left) is to predict generic location tags. BERT is used to get a distributed representation of the input for both models. The primary task is linked to the last BERT layer, while the auxiliary task is linked to a lower layer (AL). BiLSTM and CRF models are used as context encoders and tag decoders for both tasks.	23
3.1	Multi-task model overview: The architecture shows k task-specific modules (τ) attached on top of a shared <i>Transformer</i> (γ) module. The model is trained concurrently on all tasks with task-specific inputs (X).	32

4.1	Overview of the two-step training process with contrastive pre-training and task-specific fine-tuning on CrisisMMD.	41
4.2	F1 score on <i>Informative</i> task at varying training size for our proposed model. The baseline scores (at 100% training size) are shown as dashed line.	50
4.3	F1 score on <i>Humanitarian</i> task at varying training size for our proposed model. The baseline scores (at 100% training size) are shown as dashed line.	50
5.1	Distribution over the categories provided by the coders in the manually annotated set.	62
5.2	Distribution of tweets over the months.	64
5.3	Distribution over the keywords in the manually annotated set.	65
5.4	Distribution of categories in the benchmark subsets, for the Original data (left) and the Augmented data (right).	67

List of Tables

2.1	Examples of crisis tweets tagged with fine-grained location types. The subsequences representing location mentions are highlighted with pink, and their corresponding tags (in subscript) are highlighted with blue.	12
2.2	Location types and their descriptions, together with type distribution (as raw numbers and percentages) in the <i>Mixed</i> and <i>Covid</i> datasets, respectively.	18
2.3	Keywords used to collect tweets and the number of tweets from each event in the <i>Mixed</i> and <i>Covid</i> datasets, respectively.	18
2.4	Statistics for the number of tweets (Tw), tokens (Tk) and location entities (En) in the <code>train/test/dev</code> subsets of the <i>Mixed</i> and <i>Covid</i> datasets, respectively. . .	20
2.5	Hyperparameter search space for the BERT-BiLSTM-CRF model. The best overall values based on the development subset are highlighted with bold font.	25
2.6	In-domain results. Comparison between BERT-BiLSTM-CRF models and variants, and the Stanford NER baseline.	26
2.7	Cross-domain results for the BERT-BiLSTM-CRF-MTL model. Comparison between three transfer styles: zero-shot, embedding-level and model-level.	27
2.8	Error analysis tests (MFT, INV and DIR) for the capability of the model-level transfer approach to generalize the concept of a location entity.	27

2.9	Examples of location predictions for different error analysis test settings (including MFT, INV and DIR). The examples are from the <i>Mixed</i> datasets and the predictions are made using model-level transfer from <i>Covid</i> to <i>Mixed</i> . The locations are highlighted with pink. The gold labels and model predictions for entities are highlighted in blue and shown as a subscript to the corresponding locations using the convention <i>gold label</i> → <i>model prediction</i>	28
3.1	Hyperparameter configuration used in the experiments.	36
3.2	Results for Fine-grained location identification task in three settings: Single-task (ST), Location-only Multi-task (LMT) and all-multi-task (MT) with three variants of transformers (γ) as shared module. The results are grouped by dataset: NY, NZ and Combined.	37
3.3	Average (macro) performance improvement over three datasets for the three transformer variants, when comparing single-task model (ST) to location-only multi-task (LMT) and all-multi-task (MT) model.	38
4.1	Dataset sizes for informative and humanitarian tasks from <i>CrisisMMD</i> as provided in the benchmark splits by Ofli et al. (2020)	45
4.2	Distribution of pretraining dataset for two settings: <i>CrisisMMD only</i> and <i>CrisisMMD+DisRel</i>	46
4.3	Distribution of normalized cosine similarity for “ <i>Similar</i> ” and “ <i>Complementary</i> ” text-image pairs from <i>DisRel</i> pre-training development set for the three pre-trained variants of the CLIP model with mean score shown as μ	49
4.4	Results of various baselines models and our CLIP based model on Informative task.	51
4.5	Results of various baselines models and our CLIP based model on Humanitarian task.	51
5.1	Average weighted F1 score (W-F1), precision, recall and macro average F1 scores for different experiments. The results shown are the average of five runs.	69

5.2	Confusion Matrices for the model’s prediction for original and augmented benchmark datasets.	70
5.3	Examples of tweets from the minority class <i>CAG</i> that were incorrectly classified by the models.	72

Acknowledgments

First of all, I would like to express my gratitude to my advisor Professor Doina Caragea. Her guidance and support has been immense throughout this journey and has helped me become the researcher I am today. She has constantly pushed me to think critically, perform meaningful research and write good papers. I feel fortunate to have Dr. Caragea as my mentor.

I would also like to thank my PhD advisory committee members: Professor Torben Amtoft, Professor Mitchell Neilsen and Professor Raluca Cozma for their time and valuable feedback.

Next, I would like to thank my collaborators: Kyle Glandt, Maria Traskowski, Rus Refati, Dr. Sifan Xu, and Dr. Chien-fei Chen.

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

I would also like to remember my friends Anweshak Karki, Manish Khanal, Sagar Chhetri, and many more I will not be able to fit here. It has been a long and arduous journey, but my friends were always one call away to make it feel less so. I would also like to remember the friends I made here in Manhattan, KS: Rijesh Karmacharya, Jason Adhikari and Avantika Gurung who picked me up at the airport and fed me my first homely meal in this country. They and all my friends from G-18 were the first people I got to know here in town and I have never felt away from home because of them.

At last but not the least, I would like to thank my family. My grandmother, who is my biggest cheerleader, always has the biggest smile on her face when we talk over the phone. Her wisdom and blessings have had a major role in all my achievements. I would like to thank my mom and dad, who sacrificed so much so me and my brother could achieve great things in life. I would like to thank my brother, a wiser and better version of me. Finally, I would like to thank my wife, Samiksha, the love of my life. She has been the constant rock in this ride. She pushes me when I am lagging and pulls me back when I fly too close to the sun. I love you pyar!

Dedication

To my parents and my brother

For their constant love, support and blessings!

&

To my beautiful wife

For everything!

Chapter 1

Introduction

During emergent crises, there is a huge influx of information from various sources which makes the task of manually collecting and managing reliable information infeasible. Emergency response and critical resource management during such crises require time-critical and reliable information to mitigate potential health and socio-economic impact ([Irfan and Resnick, 2018](#); [Sherali et al., 2004](#)). In this work, we investigate various approaches in identifying key information relevant for emergency responders and policy makers with an over-arching goal of building a pipeline of systems that can effectively collect and identify actionable and relevant information from a noisy source of data.

However, before we build such systems, it is important to define what we mean by actionable information. [Kropczynski et al. \(2018\)](#) performed a survey including emergency help-line dispatchers, supervisors and first responders to identify “5 Ws” of actionable information. According to the study, the answers to the questions of “Where”, “When”, “Who”, “What” and “Why” represent information that is critical in making effective decisions during a crisis. The answers to these questions provide the information about location, time, affected parties and causes of the crisis. Besides these five W-questions, we can also look into the response of general public to these crises and other consequent events to gather information about public perception and stance ([Glandt et al., 2021](#)) that can aid policy-making ([Charalabidis and Loukis, 2012](#)) and help us prepare for the future.

1.1 Social media as a data source

One of the direct ways to communicate relevant information is to call the emergency help-lines. However, such emergency help-lines have been known to be overwhelmed or out of reach during crises ([Stiegler et al., 2011](#)). Beside emergency help-lines, social media and digital news media are other popular communication platforms. Social media, in particular, has evolved into a platform for people to share their concerns, report information as eyewitnesses of events, and also call for help, especially during crisis situations ([Kaewkitipong et al., 2012](#); [Wendling et al., 2013](#)). Moreover, many prior studies, e.g. [Kropczynski et al. \(2018\)](#), have shown that social media posts contain significant amount of relevant information. Therefore, the ease of access, widespread usage and presence of timely and relevant information make social media an important source of data for building tools that can aid in crisis response and management.

However, due to the unstructured nature of data shared in these digital media, inherent noise and potential misinformation ([Wang et al., 2019](#)), extraction of actionable information is a challenging task ([Wendling et al., 2013](#)). There have been active research on identifying relevant information with respect to emergency response and crisis management from the social media. [Alam et al. \(2018b\)](#) created a multimodal dataset consisting of posts made on Twitter (also known as tweets), a popular social-media platform. The tweets they collected were posted during various natural disasters to help identify informative contents and also classify them into what the authors call “humanitarian” categories. [Zahra et al. \(2020\)](#) created a similar dataset on identifying tweets made by eyewitnesses. [Chowdhury et al. \(2020\)](#) created a dataset of tweets and keyphrases that can potentially help search and classify actionable crisis related tweets.

1.2 Deep learning

Considering the challenges associated with modern data-driven emergency response and crisis management like the volume, velocity, variety and veracity of data and the time-critical nature of the task, it is clear that manual efforts fall far short of the need. Moreover, prior research has also shown that traditional machine learning approaches are not very effective in handling these challenges

(L'heureux et al., 2017). Deep learning, however, is known to be highly effective in making use of large volume of data by learning latent representations that can then be used for down-stream tasks (Sejnowski, 2020; Sun et al., 2017).

Deep-learning models have grown to be highly effective in a variety of applications, to the extent that they exceed human performance in some cases (Hekler et al., 2019; Rank et al., 2020). The approach has, consequently, been immensely popular in the last decade. Some notable milestones in the exponential growth of deep-learning approaches are the AlexNet model for image analysis (Krizhevsky et al., 2012), introduction of attention mechanism (Bahdanau et al., 2014) and transformer models (Vaswani et al., 2017) for contextual language modeling, among others.

With the models growing deeper and larger, they have become voracious consumers of huge amounts of data (Barbedo, 2018; Oyedare and Park, 2019). Furthermore, with the advent of widespread digital media, data has been more accessible than ever. However, training a model for a particular problem, generally, requires task-specific quality labelled data, which is still a prized commodity. On an abstract level, there are two main approaches to training deep-learning models: supervised learning and self-supervised learning. Depending on the problem at hand, these approaches define the need for labelled data. Supervised learning is a learning strategy where the model's objective is to learn the mapping from input data to the corresponding output. Supervised learning, hence, requires a set of input-output pairs in the form of labelled data. Building labelled data, however, may require domain expert's input and intensive human effort to perform annotation. Such annotation effort may not always be feasible, especially in time-critical problems like disaster and crisis response and management.

As the name suggests, self-supervised learning does not require manual labelling and the model learns the representation from the intrinsic relations in the data. An example of application of such learning strategy is a language model that takes in raw textual data as input and learns contextual representation of words using the context (neighbouring words) as intrinsic relations. Such language models do not require human-annotated labels and can learn semantic and syntactic relations between words without human supervision. Self-supervised learning helps address some of the issues associated with limited labelled data by making use of unlabelled data. The representations learned using the self-supervised approach can then be used to learn task-specific models on

relatively smaller labelled datasets.

In this work, we use various supervised and self-supervised deep-learning approaches, including multi-task learning, representation learning and human-in-the-loop learning, to build systems that can help with crisis response management.

1.3 Contributions of this dissertation

The trend of using larger models that consume a lot of data is also partly due to the boom of digital social media and web technologies in the last decade. The web and social media has become an integral part of modern human life. As a result, a massive influx of information is shared and collected with every passing moment on the internet. Moreover, with modern web technologies, we have come a far way in terms of modes of communication from traditional text based communication to a variety of different modalities such as images, videos, audios, documents, etc. Addressing the challenges of multimodality, noisy data source and limited labelled data, this dissertation proposes the following approaches to help crisis response management and data-driven policy and decision making.

1.3.1 Data collection

In an effort to automate data collection and management from social media, Twitter in particular, we propose a Twitter data streaming application that can run multiple Twitter data stream jobs simultaneously, based on the corresponding set of initial keywords.

The application makes use of the Twitter streaming API ¹ to gather tweets containing the defined keywords, as shown in Figure 1.1. The application uses a queue to handle the potentially large volume of tweets from multiple stream jobs. The queue processor performs initial filtering, such as English language filtering before storing the tweets into the database. The set of keywords are updated periodically using the most frequent hashtags from collected tweets to collect relevant tweets as the topic of discussion evolves over time.

¹<https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/api-reference/post-statuses-filter>

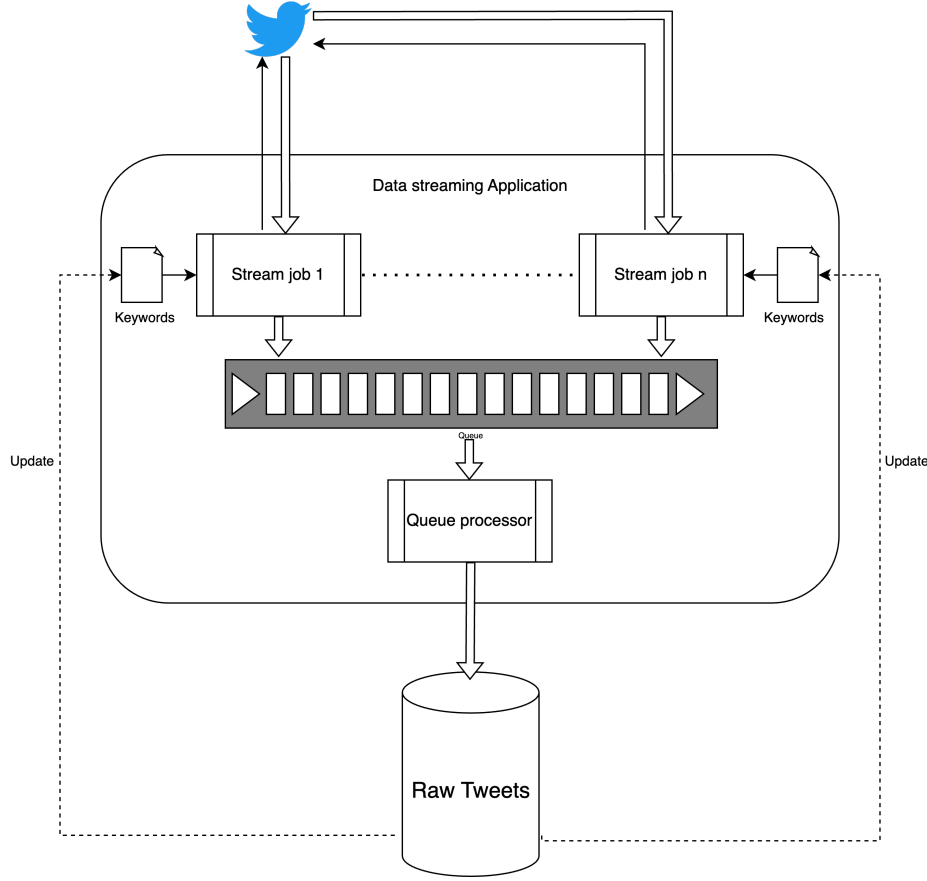


Figure 1.1: *Proposed Twitter data streaming pipeline.*

Using this application, we collected over 3 TB of raw tweets about COVID-19 and the events surrounding the pandemic. Besides that, we also used the application to collect tweets for various other events such as Sri Lanka Bombing², Hurricane Michael³ and Hurricane Florence⁴.

1.3.2 Identification of fine-grained location mentions in crisis tweets

Identification of fine-grained location mentions in crisis tweets is central in transforming situational awareness information extracted from social media into actionable information. Prior research (Kropczynski et al., 2018; Vieweg et al., 2010) has shown that location is one of the most important piece of information for emergency response. Most prior works, however, have focused on identifying generic locations, without considering their specific types (Goyal et al., 2018; Hoang et al.,

²https://en.wikipedia.org/wiki/2019_Sri_Lanka_Easter_bombings

³https://en.wikipedia.org/wiki/Hurricane_Michael

⁴https://en.wikipedia.org/wiki/Hurricane_Florence

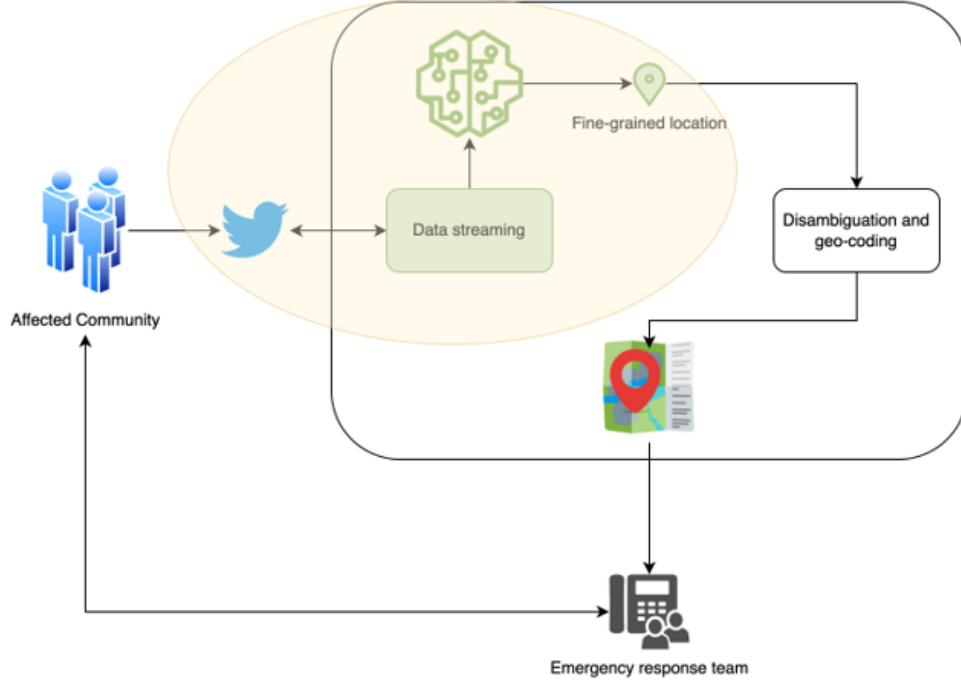


Figure 1.2: Overview of the overarching goal for the identification of fine-grained location mentions in crisis tweets. The scope of the project is highlighted in yellow oval.

2017; Kumar and Singh, 2019; Li et al., 2020). However, finding locations within a text with their associated fine-grained types can help in geo-coding and disambiguation. To facilitate progress on the fine-grained location identification task, we assembled two tweet crisis datasets and manually annotated them with specific location types. The first dataset contains tweets from a mixed set of crisis events, while the second dataset contains tweets from the global COVID-19 pandemic. In this chapter, we investigated the performance of state-of-the-art deep learning models for sequence tagging, in both in-domain and cross-domain settings using the two assembled datasets. In particular, we compared different embedding strategies such as word and character-based embeddings (Ma and Hovy, 2016; Pennington et al., 2014), transformer-based embeddings (Devlin et al., 2018) and entity-level embeddings (Yamada et al., 2020) to learn representations of input sequences and tested the model’s capability to generalize the learned representations across the two datasets. We also proposed a multi-tasking approach where the task of identifying candidate location entities without fine-grained types is used as an auxiliary task for the main task of identifying fine-grained location mentions.

The overarching goal of this project, as shown in Figure 1.2, is to build a pipeline of systems

that can collect relevant tweets from affected individuals, potentially in real-time, and identify fine-grained location mentions within the relevant tweets. The identified location mentions can then be geo-coded and visualized on the map and in turn this can aid emergency response teams in identifying hot-spots.

1.3.3 Multi-task learning to enable location mention identification in the early hours of a crisis event

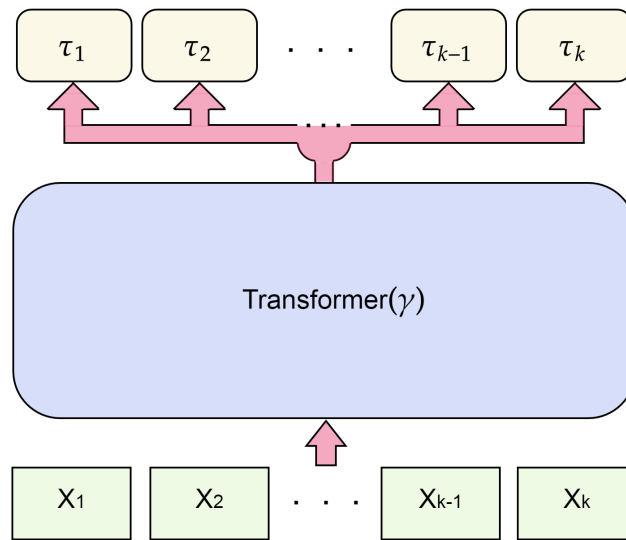


Figure 1.3: Multi-task learning architecture containing a shared transformer module (γ) and k task-specific modules (τ). $X_i (1 \leq i \leq k)$ are the datasets used for each task, respectively.

Training a robust and reliable deep learning model requires a large amount of data. In the crisis domain, building deep learning models to identify actionable information from the huge influx of data posted by eyewitnesses of crisis events on social media, in a time-critical manner, is central for fast response and relief operations. However, building a large, annotated dataset to train deep learning models is not always feasible in a crisis situation. In this chapter, we investigated a multi-task learning approach to concurrently leverage available annotated data for several related tasks from the crisis domain to improve the performance on a main task with limited annotated data as shown in Figure 1.3. Specifically, we focused on using multi-task learning to improve the performance on the task of identifying location mentions in crisis tweets when using additional

crisis-related tasks such as eyewitness-account classification, humanitarian categories classification and keyphrase identification. In particular, we investigated the effect of task similarity in multi-task learning along with the effect of the size of the embedding module in the overall performance gain when using multiple tasks. Our results showed that when a small amount of data is available, the model may benefit more from using other related tasks rather than using additional data for the same type of task (Khanal and Caragea, 2021).

1.3.4 Contrastive learning for multimodal classification of crisis related tweets

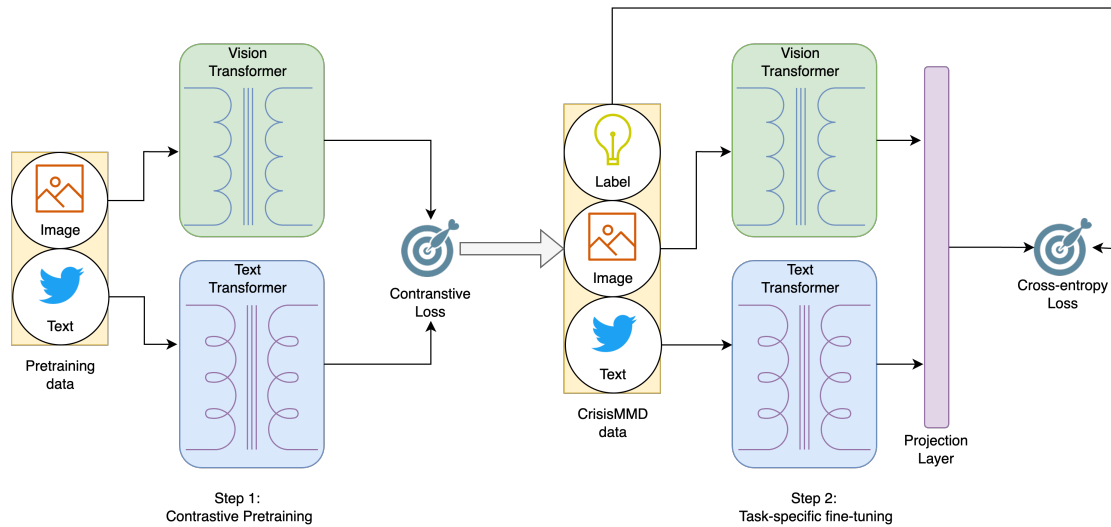


Figure 1.4: Overview of the two-step multimodal training process with contrastive pre-training and task-specific fine-tuning.

Multimodal tasks require learning a joint representation of constituent modalities of data. Several existing approaches use separate embedding modules for each modality and jointly train them on downstream tasks using fusion layers. Contrastive learning, however, aims to learn a joint representation by contrasting each image-caption pair with other pairs. One key assumption in this approach is that the caption describes the content of the image. However, this does not hold true for tweets. For crisis related tweets, prior research has shown that the image-text relation is not always complementary. The parity of informativeness may differ between the text and associated

image within the tweet. Using the CrisisMMD dataset (Alam et al., 2018b), a publicly available dataset, we showed that domain-specific pre-training using contrastive learning outperforms existing approaches for multimodal classification of crisis related tweets. The overview of the two step training process is shown in Figure 1.4.

1.3.5 Using content analysis and machine learning to identify COVID-19 information relevant to low-income households on social media

COVID-19 pandemic has caused great distress in the lives of many populations. Low-income households are among the most severely impacted groups in the United States and across the globe. Using social media, this chapter aims to identify and organize the information about the impact of the pandemic on low-income households. We used content analysis to derive an annotation protocol and manually annotated a tweet dataset using this protocol. Furthermore, we used machine learning to learn models from the annotated dataset. We also employed a human-in-the-loop data augmentation procedure to improve the model’s performance for the underrepresented classes. Our results showed that using carefully annotated data, automated machine learning models can be trained and employed to identify information relevant to low income households, potentially in real time (Khanal et al., 2021).

1.3.6 Summary

In summary, the contribution of this dissertation includes various approaches focused on identifying actionable information from social media and other digital contents that can help in crisis response management and data-driven decision making. Some of the contributions of this work have already been published in peer-reviewed conferences as shown below:

- S. Khanal, M. Traskowsky, and D. Caragea. Identification of fine-grained location mentions in crisis tweets. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 7164–7173, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.776>

- S. Khanal and D. Caragea. Multi-task learning to enable location mention identification in the early hours of a crisis event. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4051–4056, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.340. URL <https://aclanthology.org/2021.findings-emnlp.340>
- S. Khanal, R. Refati, K. Glandt, D. Caragea, S. Xu, and C.-f. Chen. Using content analysis and machine learning to identify covid-19 information relevant to low-income households on social media. In 2021 IEEE Intl Conf on Parallel Distributed Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking (ISPA/ BDCloud/ SocialCom/ SustainCom), pages 1522–1531, 2021. doi: 10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00205

Furthermore, Chapter 1.3.4, focused on “Contrastive learning for multimodal classification of crisis related tweets” will also be submitted as a publication to a peer-reviewed conference.

Besides these publications that are directly included in the dissertation, the following publications are additional contributions that I have made during my PhD studies. Although they have not been included in the dissertation, they are pertinent to the use of deep learning and digital media in crisis management.

- K. Glandt, S. Khanal, Y. Li, D. Caragea, and C. Caragea. Stance detection in covid-19 tweets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1596–1611, 2021
- N. M. Nafi, A. Bose, S. Khanal, D. Caragea, and W. H. Hsu. Abstractive text summarization of disaster-related document. In ISCRAM 2020 Conference Proceedings–17th International Conference on Information Systems for Crisis Response and Management, 2020

Chapter 2

Identification of fine-grained location mentions in crisis tweets

2.1 Introduction

Affected individuals during natural disasters and other crisis situation often turn to social media (e.g., Twitter or Facebook) to report useful information, or ask for help ([King, 2018](#); [Sakaki et al., 2010](#); [Vieweg et al., 2010](#)). Information contributed on social media by people on the ground can be invaluable to emergency response organizations in terms of gaining situational awareness, prioritizing resources to best assist the affected population, addressing concerns, and even saving lives ([King, 2018](#)).

Many recent studies have focused on identifying informative tweets posted by individuals affected by a crisis, and classifying those tweets according to situational awareness categories useful for crisis response and management ([Imran et al., 2015b](#)). However, for situational awareness information extracted from social media to be actionable, knowing the corresponding geographic location is of key importance. Unfortunately, only a very small percentage of tweets are geotagged ([Mahmud et al., 2012](#)). Furthermore, even when geolocation information is available, that location may not be the location mentioned in the tweet text ([Ikawa et al., 2013](#)). According to [Vieweg et al. \(2010\)](#), the location in the tweet text is usually the location needed for monitoring and/or

No.	Tweet text
1	Roads in Calhoun County _{ctc} are underwater, access to the Port Lavaca Causeway _{lan} is flooded, the bridge is closed.
2	Very extensive damage sustained throughout Wilmington _{ctc} NC _{sta} from Hurricane Florence.
3	Big tree fell on power lines and blocking Brown Ave _{oth} near Washington St _{oth} in Orlando's _{sta} Thornton Park _{lan}
4	There are now more confirmed cases of coronavirus in New York City _{ctc} than there are in all of South Korea _{con}
5	South Asia _{reg} is quickly marching towards being the new epicenter of covid 19
6	The difference in COVID 19 cases and deaths between New York _{sta} and California _{sta} continues to be astounding

Table 2.1: Examples of crisis tweets tagged with fine-grained location types. The subsequences representing location mentions are highlighted with pink, and their corresponding tags (in subscript) are highlighted with blue.

responding to an emergency. Table 2.1 shows several examples of tweets posted during recent hurricanes (first three tweets) and during the Covid-19 crisis (last three tweets). As can be seen, locations are mentioned at different levels of granularity, from region and landmark to city, state and country. Furthermore, the same location name, in our Covid-19 examples - *New York*, can be associated with different location types, such as *city* (tweet 5) and *state* (tweet 6). Information about the tags of the ambiguous entities can be used to disambiguate the corresponding locations and link them to physical locations. Therefore, tools for identifying fine-grained location types directly from the texts of crisis tweets are greatly needed.

Location identification has been frequently addressed as part of the broader named entity recognition (NER) task (Goyal et al., 2018; Li et al., 2020). Some studies have focused specifically on the task of identifying generic location mentions (without considering the type of location) in tweet text (Hoang et al., 2017), and even disaster tweet text (Kumar and Singh, 2019). Other studies have focused on identifying fine-grained point-of-interest (POI) type locations, useful for location-based services (Ji et al., 2016; Li and Sun, 2014; Malmasi and Dras, 2015; Xu et al., 2019).

To the best of our knowledge, there are no publicly available, manually annotated datasets that can facilitate progress on the task of identifying fine-grained locations (including, city, state, country,

region, landmark) in crisis tweets, despite the benefits provided by the use of social media data in monitoring and responding to a crisis. To address this need, we have assembled two datasets for identifying fine-grained locations in crisis tweets. The first dataset, called *Mixed*, consists of tweets crawled during five crisis events, specifically, Nepal Earthquake, Queensland Floods, Srilanka Bombing, Hurricane Michael and Hurricane Florence. The second dataset, called *Covid*, consists of a random set of coronavirus-related tweets crawled between February 27 and April 7, 2020. We used Amazon Mechanical Turk (AMT)¹ to annotate the datasets using six location types (shown in Table 2.2).

Given the success of deep learning approaches for NER tasks (Li et al., 2020), we propose to use a sequence labeling deep learning approach to identify fine-grained locations in crisis tweets. More specifically, we use an architecture considered state-of-the-art for NER (Li et al., 2020), which has three main components: a *distributed representation* model for input, a *context encoder* and a *tag decoder*. We use a BERT model (short for Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) as a distributed representation of the input, a BiLSTM model (short for Bidirectional Long Short-Term Memory) (Hochreiter and Schmidhuber, 1997) as a context encoder, and a CRF model (short for Conditional Random Fields) (Lafferty et al., 2001) as a tag decoder. An overview of the model architecture is shown in Figure 2.1. We study the combined BERT-BiLSTM-CRF model in both in-domain and cross-domain settings to gain insights into its effectiveness, as well as its transferability to a new emerging crisis for which labeled data may not yet be available. In summary, the contributions of this work are as follows:

- We create two datasets of tweets from a mixed set of crisis events and from the Covid-19 crisis, respectively. The tweets are manually annotated with fine-grained location types, including city, state, country, region, landmark.
- We propose to identify fine-grained locations in crisis tweets using state-of-the-art BERT-BiLSTM-CRF models within a multi-task learning framework (where fine-grained location identification is the main task and generic location identification is the auxiliary task).

¹<https://www.mturk.com/>

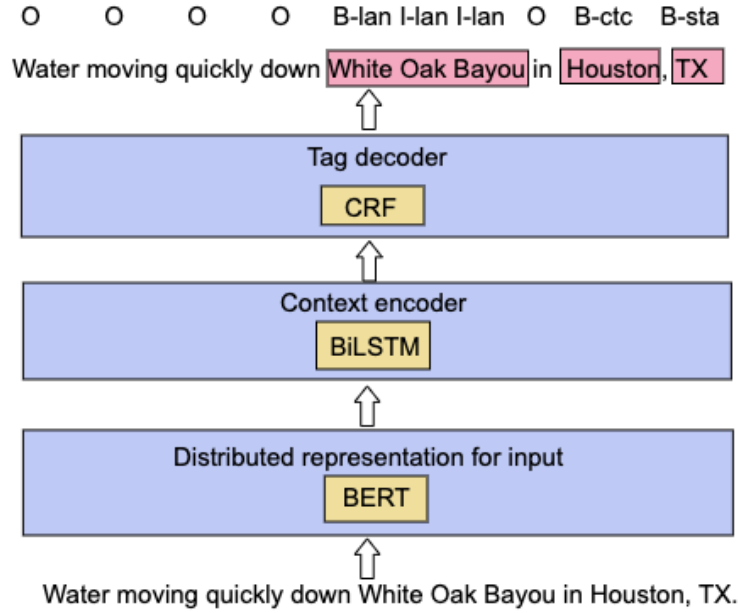


Figure 2.1: Overview of the the deep learning architecture for fine-grained location identification.

- We perform extensive experiments on the *Mixed* and *Covid* datasets, respectively, in both in-domain and cross-domain settings.

2.2 Related work

2.2.1 NER and location mention identification

NER is a well-researched problem in natural language processing (NLP) (Goyal et al., 2018; Li et al., 2020). Text-based location identification has been traditionally addressed as part of the broader NER task, although some works focus specifically on location identification (Han et al., 2014; Kumar and Singh, 2019; Lingad et al., 2013; Magge et al., 2019). Most of the works that identify locations simply tag location mentions, as opposed to identifying fine-grained location types (Li et al., 2020). For example, Lingad et al. (2013) aim to identify mentions of locations (including geographic locations and points of interest) in disaster tweets, by using standard NER taggers (pre-trained or retrained), and report best performance using retrained Stanford NER (Finkel et al., 2005). Also in the context of emergencies, Kumar and Singh (2019) use a convolutional

neural network (CNN) approach to identify location references in crisis tweets, regardless of their specific types.

2.2.2 Fine-grained location types

Some recent works have considered fine-grained location types, such as city, state, country (Anand et al., 2017; Inkpen et al., 2015; Lal et al., 2019; Qazi et al., 2020). While focused on Covid-19 tweets, Qazi et al. (2020) use a gazetteer approach to infer the geolocation of tweets, based on user and tweet information. Closest to our goal of identifying fine-grained locations in disaster tweet texts, Inkpen et al. (2015) propose a CRF-based approach to identify countries, states/provinces and cities using a Twitter dataset annotated according to guidelines provided in (Mani et al., 2010). They make use of hand-crafted features, including gazetteer features, to train a CRF model. As opposed to (Inkpen et al., 2015), we use a larger set of location types, and approaches that preclude the need for manually crafted features and gazetteers.

Other works on fine-grained location focus on identifying point of interests locations, such as restaurants, hotels, parks, etc. and linking them to pre-defined location profiles (Ji et al., 2016; Li and Sun, 2014; Xu et al., 2019). Li and Sun (2014) build a POI inventory (which can be seen as a noisy version of a gazetteer), and a time-aware POI tagger. The time-aware POI tagger is a CRF trained to extract and disambiguate fine-grained POIs. Ji et al. (2016) extend the POI tagger in Li and Sun (2014) by proposing a joint framework that achieves POI recognition and linking to pre-defined POI profiles simultaneously. Xu et al. (2019) address the same problem of identifying fine-grained POIs and linking them to location profiles. However, they use a deep learning model (specifically, BiLSTM-CRF) to avoid the need for manually designed features, and subsequently use a collection of location profiles to perform the linking. The definition of fine-grained POI tagging is different from our definition of fine-grained location tagging, in the sense that we aim to assign specific types/tags to location entities, as opposed to identifying generic (yes/no) POI tags, and then linking the tags to pre-defined profiles, as in prior works (Ji et al., 2016; Li and Sun, 2014; Xu et al., 2019). Moreover, we want to avoid the use of gazetteers to ensure that the models are resilient to the informal nature of the language used in tweets. Similar to (Xu et al., 2019), we also want to

avoid the need for manually designed features, and thus focus on deep learning approaches.

2.2.3 State-of-the-art approaches for NER

State-of-the-art approaches for NER, in general, and location identification, in particular, are sequence labeling type approaches based on deep learning language models (Li et al., 2020). More specifically, competitive architectures consist of three components: distributed representations of the input, a context encoder model, and a tag decoder model. Both character-level and word-level embeddings (or their combination) have been used to represent the NER input in recent works (Goyal et al., 2018), with BERT contextual embeddings being among the most successful (Li et al., 2020). In terms of context encoders and tag decoders, recurrent neural networks, most often, BiLSTM networks, and CRF, respectively, contribute to some of the best results on benchmark NER datasets (Baevski et al., 2019; Jiang et al., 2019; Liu et al., 2019b; Luo et al., 2019). Given the state-of-the-art architectures for the NER task, our proposed model consists of three components: BERT, BiLSTM and CRF, for the input representation, context encoder and tag decoder, respectively.

2.2.4 Location datasets

Most previous works on location identification in tweet texts are focused on general tweets (Inkpen et al., 2015; Liu et al., 2014) with a few notable exceptions of works focused on crisis tweets (Kumar and Singh, 2019; Lingad et al., 2013; Qazi et al., 2020). However, the datasets used in these works are not all available (Kumar and Singh, 2019; Lingad et al., 2013). Even when available, the datasets focus on identifying location mentions without specifically identifying the fine-grained type of the location mentions (Liu et al., 2014). Qazi et al. (2020) used a gazetteer-only approach to annotate tweets with geolocations, and the resulting annotations are not very accurate. Inkpen et al. (2015) used gazetteer, as well as manual annotations in their dataset, however, that dataset is not focused on crisis tweets.

2.3 Datasets

One main contribution of our work is to construct two benchmark datasets for identifying fine-grained locations (see Table 2.2) useful for crisis monitoring and response. The datasets cover events that are different in nature, to enable studies in both in-domain and cross-domain settings.

2.3.1 Data collection

The first dataset, called *Mixed*, contains tweets posted during four natural disasters and one man-made disaster that happened in specific geographical regions. The second dataset, called *Covid*, contains tweets posted during the Covid-19 pandemic, and thus has worldwide coverage. More specifically, the tweets in the *Mixed* dataset were crawled during the following events: Nepal Earthquake, Queensland Floods, Srilanka Bombing, Hurricane Michael and Hurricane Florence. The tweets from Nepal Earthquake and Queensland Floods were obtained from (Alam et al., 2018a). Tweets from Srilanka Bombing, Hurricane Michael and Hurricane Florence were crawled locally using the Twitter streaming API, and a random sample of unique English tweets was included in the *Mixed* dataset. Tweets from Covid-19 pandemic were also crawled locally from February 27th to and April 7th, 2020. A random sample of unique English tweets was included in the *Covid* dataset. The keywords used to crawl the tweets for each event are listed in Table 2.3.

Furthermore, to pre-train the BERT model for the *Mixed* dataset, we collected a larger set of tweets pertaining to various crisis events from prior works (Alam et al., 2018a,b; Imran et al., 2016; Nguyen et al., 2017; Olteanu et al., 2014, 2015b) in addition to the locally crawled tweets. For the *Covid* dataset, however, we only used the locally crawled tweets to pre-train the BERT model.

2.3.2 Data annotation and quality assessment

To prepare the tweets for annotation, the following pre-processing was performed. User mentions were anonymized by replacing them with a generic `user` keyword, and links were removed from the tweet text. Special characters, including `!=#$%^&*()+[]{};\':"|<>?`, and non-printable ASCII characters were also removed. The tweet text was tokenized to enforce annotation at the

Type	Descr.	Mixed Distr.		Covid Distr.	
		No.	%	No.	%
con	Country	1,763	28.31	1,819	53.06
sta	State	797	12.80	518	15.11
reg	Region, Continent	1,242	19.95	396	11.56
ctc	City, Town, County	764	12.27	158	4.61
lan	Building, Landmark	1,190	19.11	391	11.42
oth	Other	471	7.56	146	4.24
All	Entities	6,227	100.00	3,428	100.00

Table 2.2: Location types and their descriptions, together with type distribution (as raw numbers and percentages) in the Mixed and Covid datasets, respectively.

Dataset	Event	Keywords	Size
Mixed	Nepal Earthquake and Queensland Floods (Alam et al., 2018a)	N/A	167
	Srilanka Bombing (ours)	Sri lanka attack, Sri lanka terror, Sri Lanka horror, Sri Lanka easter	1171
	Hurricane Michael and Hurricane Florence (ours)	hurricane michael, hurricanemichael, hurricane florence, hurricaneflorence	2758
Covid	Covid-19 (ours)	#coronavirus, corona virus, #Coronavid19, #coronavirususa, #covid19, #covid-19, #coronapocalypse, #quarantinelifelife, #socialdistancing	5243

Table 2.3: Keywords used to collect tweets and the number of tweets from each event in the Mixed and Covid datasets, respectively.

token level and avoid accidental annotation of token fragments. Tweet tokens were annotated with six location types using the BIO scheme (where B stands for Beginning, I stands for Inside and O stands for Outside of a location entity). The location types together with their brief descriptions are shown in Table 2.2. Examples of annotated tweets are shown in Table 2.1, where the first three tweets are representative of the *Mixed* dataset, and the last three are representative *Covid*.

We used feedback from a local annotator to iteratively develop and improve a custom annotation tool for our task. The tool was subsequently deployed to Amazon Mechanical Turk (AMT).

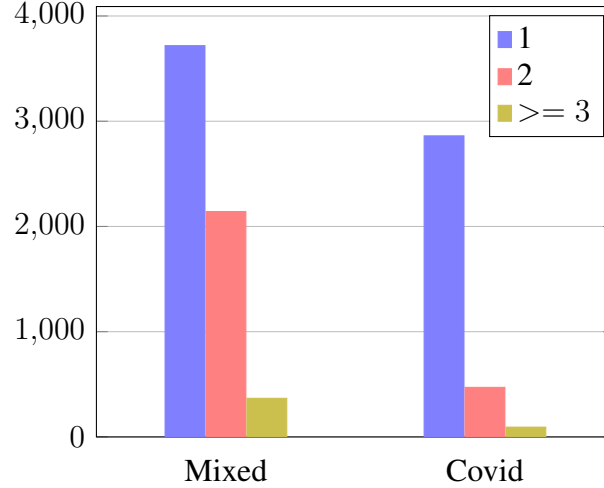


Figure 2.2: Entity distribution by number of tokens in a location in the *Mixed* and *Covid* datasets, respectively.

Annotators were provided with definitions of the location types included in our study, together with precise instructions for annotation, and examples of annotated tweets, such as those in Table 2.1. Each tweet was annotated by at least 3 annotators. Only entities where two or more annotators agreed were included in the datasets. We used the `nltk.metrics` package² to calculate inter-annotator agreement metrics (Artstein and Poesio, 2008), specifically, Cohen’s Kappa scores (Cohen, 1960). The Cohen’s Kappa scores that we obtained for inter-annotator agreement were 0.63 and 0.62 and the average pairwise F1-scores for inter-annotator agreement were 68.87 and 65.86 for the *Mixed* and *Covid* datasets, respectively. According to Cohen (1960), these scores represent substantial agreement. The final number of location entities in the *Mixed* dataset is 6,227, while the number of location entities in the *Covid* dataset is 3,428. The distributions of the location entities over the six location types included in our study are shown in Table 2.2, and the entity distributions by the number of tokens in a location are shown in Figure 2.2.

2.3.3 Benchmark datasets

To enable progress on fine-grained location identification in crisis tweets, and facilitate comparisons between models developed for this task (in-domain and cross-domain), we created benchmark

²https://www.nltk.org/_modules/nltk/metrics/agreement.html

datasets by randomly splitting our *Mixed* and *Covid* datasets, based on tweets, into training (`train`), development (`dev`) and test (`test`) subsets, respectively. We use the training subset to train our models, the development subset to select hyperparameters and the test subset to evaluate the final performance of the models. Statistics for the *Mixed* and *Covid* datasets in terms of number of tweets, tokens, entities in the `train`, `test` and `dev` subsets, respectively, are shown in Table 2.4. The benchmark datasets is available publicly ³.

Data	No.	Train	Test	Dev	Total
Mixed	Tw	2,620	820	656	4,096
	Tk	73,622	23,253	18,511	115,386
	En	4,001	1,237	989	6,227
Covid	Tw	3,355	1,049	839	5,243
	Tk	103,646	32,674	25,798	162,118
	En	2,206	656	566	3,428

Table 2.4: Statistics for the number of tweets (*Tw*), tokens (*Tk*) and location entities (*En*) in the `train/test/dev` subsets of the *Mixed* and *Covid* datasets, respectively.

2.4 Background and approaches

The task of identifying fine-grained locations in tweet text can be formulated as follows: Given a set of (X, Y) pairs, where X is a text sequence/tweet $X = \{x_1, \dots, x_n\}$ with n tokens, and Y is a tag sequence $Y = \{y_1, \dots, y_n\}$ with n location tags/types (in BIO format) corresponding to the tokens in the sequence X ; our sequence tagging task is to find a mapping $f_\theta : \mathbf{X} \rightarrow \mathbf{Y}$ (with parameters θ) from input sequences to fine-grained locations.

We use a state-of-the-art deep learning architecture, shown in Figure 2.3, to address the fine-grained location identification task. The architecture uses a distributed representation that consists of BERT embeddings, a BiLSTM network as the context encoder, and a CRF as the tag decoder. It also uses multi-task learning to include an auxiliary task. We will explain how we used some of these components below. Furthermore, we will discuss the main approach and several variants that we studied.

³<https://github.com/sarthakksu/finegrained-location-data>

2.4.1 BERT embeddings

BERT (Devlin et al., 2018) is a language model that uses a deep bidirectional self-attention transformer encoder architecture (Vaswani et al., 2017) to encode sentences and their tokens. A generic model is pre-trained on a large corpus of un-annotated text (e.g., Wikipedia) using two self-supervised learning tasks: masked word prediction (a.k.a., masked language modeling, or MLM) and next sentence prediction (NSP). BERT takes as input a sequence of word tokens, where the first token is a special token denoted by [CLS]. The multi-layer bidirectional transformer encoder transforms the original input embeddings into a distributed representation. A generic BERT model, pre-trained on a large corpus, can be further *pre-trained* and/or *fine-tuned* for specific NLP tasks (Devlin et al., 2018). To customize BERT to crisis tweets, we start with a pre-trained model, `bert-base-uncased`, and further pre-train it on a large collection of mixed tweets and Covid-19 tweets, respectively, using the masked language modeling (MLM) task. We use the Huggingface library (Wolf et al., 2019) to pre-train BERT. The pre-trained BERT embeddings are provided as input to a BiLSTM, whose output is further provided as input to a CRF model, which identifies locations.

BERT uses WordPiece (Wu et al., 2016) tokenization, which can break a single word into multiple pieces. This presents a challenge in our input representation, where the tokenized input has a different token collections than the tagged output. To account for this, when an input token is broken down into pieces by the WordPiece tokenizer, we assign the original tag to the first piece of the token and label the remaining pieces with a special token X masked for the final prediction. This approach preserves the token-label mapping of the input sequence and aligns with the tokenization requirement of BERT at the same time.

2.4.2 BiLSTM

LSTM networks (Hochreiter and Schmidhuber, 1997) are a type of recurrent neural networks (RNN) that can be used to capture dependencies in sequence data, including long term dependencies. At the core of an RNN network, including LSTM networks, there is a recurrent cell, which enables the network to pass information from one time step to the next one. A standard LSTM has one layer

corresponding to one cell, and processes the sequence in the forward direction. A bidirectional LSTM (BiLSTM) includes a second layer/cell, which processes the sequence in reverse direction. We use a BiLSTM network as context encoder to capture sequence dependencies in both forward and reverse directions.

2.4.3 Conditional random fields (CRFs)

CRFs ([Lafferty et al., 2001](#)) are commonly used as decoders in sequence labeling tasks. A CRF model takes as input a distributed representation $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of the tokens in a sequence $X = \{x_1, \dots, x_n\}$, and outputs a decoding $\mathbf{y}$ of the input representation $\mathbf{X}$. Context-sensitive representations produced by BiLSTM models are frequently used as input representations for CRF models. A CRF model is trained to maximize the log-probability of the correct sequence of tags as defined below (or equivalently to minimize the loss given by the negative log-probability):

$$\log(p(\mathbf{y}|\mathbf{X})) = s(\mathbf{X}, \mathbf{y}) - \log\left(\sum_{\tilde{\mathbf{y}} \in \mathbf{Y}_{\mathbf{x}}} e^{s(\mathbf{X}, \tilde{\mathbf{y}})}\right) \quad (2.1)$$

where $s(\mathbf{X}, \mathbf{y})$ is a tagging score function, and $\mathbf{Y}_{\mathbf{x}}$ represents the set of all possible sequences of tags for the input sequence $\mathbf{X}$. During inference, a sequence $\mathbf{X}$ is decoded into the tag sequence $\mathbf{y}^*$, which maximizes the tagging score below:

$$\mathbf{y}^* = \arg \max_{\tilde{\mathbf{y}} \in \mathbf{Y}_{\mathbf{x}}} s(\mathbf{X}, \tilde{\mathbf{y}}) \quad (2.2)$$

We use a CRF as our tag decoder in the generic model architecture shown in [Figure 2.1](#).

2.4.4 Multi-task learning

Multi-task learning (MTL) ([Caruana, 1997](#)) is focused on training multiple related tasks in parallel to leverage shared patterns and improve the performance of each individual task. Generally, in the MTL framework, the related tasks share some lower layers that extract common features, while they also have task-specific layers. In this work, we use the task of tagging generic location mentions, as

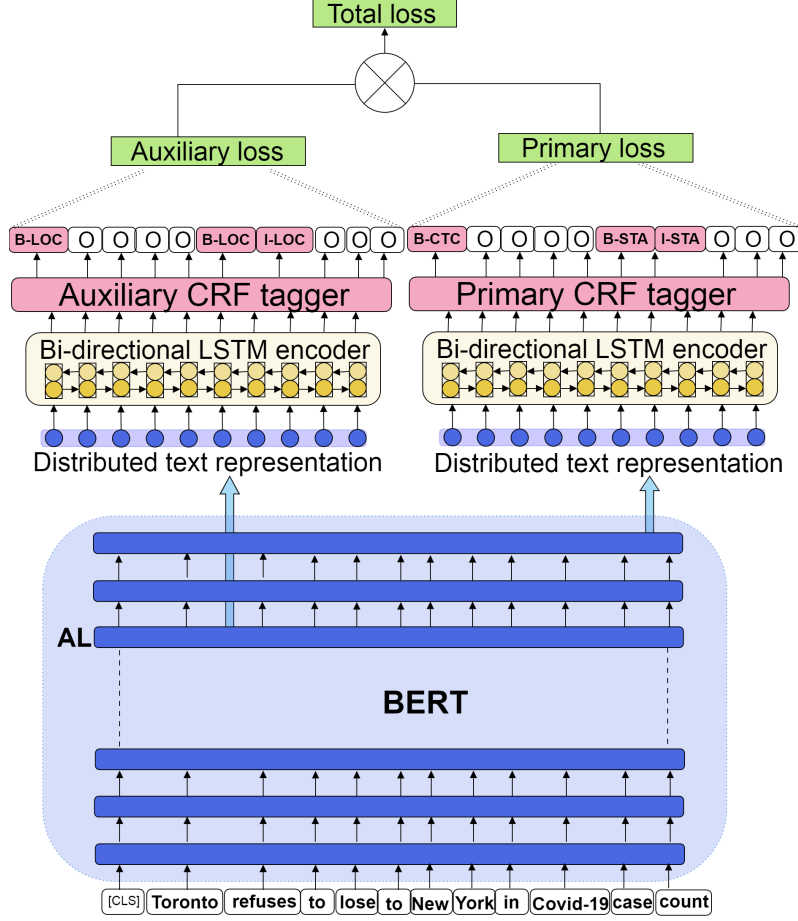


Figure 2.3: *Proposed model architecture. The model can be seen as an MTL model, with two objective corresponding to two tasks. The primary task (right) is to predict fine-grained location tags, while the auxiliary task (left) is to predict generic location tags. BERT is used to get a distributed representation of the input for both models. The primary task is linked to the last BERT layer, while the auxiliary task is linked to a lower layer (AL). BiLSTM and CRF models are used as context encoders and tag decoders for both tasks.*

an auxiliary task for the main fine-grained location tagging task.

2.4.5 Proposed model and variants

Our proposed model, shown in Figure 2.3, can be seen as an MTL model with two objectives. The primary objective corresponds to the task of tagging the input sequence with fine-grained location types (right). The auxiliary objective corresponds to the task of tagging the input sequence with generic location/not-location tags (left). The final objective is a weighted sum of the primary and auxiliary objectives. According to [Hoang et al. \(2017\)](#), information about the presence of

locations improves the performance of the location identification task. Similarly, our hypothesis is that information about generic location mentions may improve the performance of the fine-grained location identification.

As Figure 2.3 shows, the primary task is linked to the last layer of the BERT model, while the auxiliary task is linked to a lower layer, denoted AL (auxiliary layer) in the figure. The assumption is that a lower BERT layer captures more generic features shared by the primary and auxiliary tasks. The context encoding and the tag decoding for each task are done with task-specific BiLSTM and CRF models, respectively, which are co-trained in parallel. In what follows, we denote this model by BERT-BiLSTM-CRF-MTL. To understand the contribution of the auxiliary task, we also experiment with a model similar to the model in Figure 2.3, which does not include the generic location tagging task (left). This model is denoted by BERT-BiLSTM-CRF in what follows.

2.5 Experimental setup

2.5.1 Baseline

Stanford NER (Finkel et al., 2005) uses an arbitrary order linear chain CRF model over a set of predefined word and character level features extracted from the input. The model has been used as a strong baseline for many NER models. We retrain the model with both *Mixed* and *Covid* datasets, respectively, to learn fine-grained location types.

2.5.2 Metrics

We use standard metrics, including precision (Pr), recall (Re) and F1-measure (F1), to evaluate the performance of the models trained.

2.5.3 Implementation details

The hyperparameter search-space explored in our experiments is shown in Table 2.5. We performed a grid-search with 50 trials and used the development subsets to identify overall good hyperparameter

Hyperparameters	Search space
Fine-tuned BERT layers	9, 10, 11
Auxiliary task layer (AL)	None, 6, 7, 8 , 9, 10
Hidden size of BiLSTM	64, 128, 256, 512

Table 2.5: Hyperparameter search space for the BERT-BiLSTM-CRF model. The best overall values based on the development subset are highlighted with bold font.

values (highlighted in bold font in Table 2.5). We used those values consistently in the remaining experiments. We used Glorot uniform initializer (Glorot and Bengio, 2010) to initialize the model weights, and the optimization was performed with the AdamW optimizer (Loshchilov and Hutter, 2019), with a learning rate of $1e-3$, weight decay of $1e-2$, and gradient clipping with max norm of 5. We used a dropout of 0.5, mini-batch size of 32, and a weight of 0.2 for the auxiliary loss in all the experiments. The CNN layer’s kernel size was set to 3 with 30 filters. We set a patience of 5 epochs on the development F1-measure, as early stopping of training. All experiments are run on NVIDIA Tesla V100 GPU.

2.5.4 Experiments

We conducted experiments in two settings, *in-domain* and *cross-domain*. In the **in-domain setting**, models were trained and tested on the same dataset (e.g., models were trained on Mixed-train, tuned on Mixed-dev, and tested on Mixed-test). The goal was to study: 1) the performance of the deep learning models by comparison with the traditional Standrod NER model; 2) the effect of the auxiliary task in the MTL framework. In the **cross-domain setting**, we used the best in-domain model to investigate several ways to perform transfer of information between domains: 1) a *zero-shot* transfer setting, where models trained on one dataset were tested on the other dataset (e.g., models trained on Mixed-train, tuned on Mixed-dev and tested on Covid-test); 2) an *embedding-level* transfer, where the BERT block fine-tuned on one dataset (e.g., Mixed) was used as a starting point for the BERT block of the model trained/tuned/tested on the other dataset (e.g., Covid); 3) a *model-level* transfer, where the model trained/tuned on a dataset (e.g., Mixed-train, Mixed-dev) is used as the starting point of the model for the other dataset (e.g., Covid-train, Covid-dev, Covid-test, respectively).

Dataset	Mixed		
Model	Pr	Re	F1
Stanford NLP (retrained)	82.52	65.64	73.12
BERT-BiLSTM-CRF	74.98	70.07	74.52
BERT-BiLSTM-CRF-MTL	74.58	74.75	74.67
Dataset	Covid		
Model	Pr	Re	F1
Stanford NLP (retrained)	85.71	57.62	68.92
BERT-BiLSTM-CRF	73.41	70.74	72.05
BERT-BiLSTM-CRF-MTL	77.06	69.43	73.04

Table 2.6: In-domain results. Comparison between BERT-BiLSTM-CRF models and variants, and the Stanford NER baseline.

2.6 Results and discussion

2.6.1 In-domain setting

Table 2.6 shows the in-domain results of the proposed BERT-BiLSTM-CRF-MTL model, its variants and the Stanford NER baseline model. As can be seen in Table 2.6, the proposed model is the best overall in terms of F1-measure for both *Mixed* and *Covid* datasets, with a relatively high recall compared to the baseline. Specifically, the F1-measure is 74.67% for the *Mixed* dataset and 73.04% for the *Covid* dataset. While the baseline has the highest precision overall, we argue that in the context of disaster monitoring and response, recall is more important than precision, as the final results will be reviewed by humans before any action is taken. Comparing the results for the *Mixed* and *Covid* datasets, we can see that the models have slightly better performance on the *Mixed* dataset. While this dataset contains a variety of crisis events, the events are relatively localized to specific geographical regions, which may make it easier for the models to identify the locations. As opposed to that, the *Covid* dataset has a big variety of locations as it covers a global pandemic. Nevertheless, the F1 score of the proposed model on *Covid* is 3.1% higher than the score of the baseline model, which uses manually designed features for training. Finally, when comparing the BERT-BiLSTM-CRF-MTL model (with auxiliary task) to its BERT-BiLSTM-CRF variant (without the auxiliary task), the results show that the auxiliary task can help improve the F1-measure, especially in the case of the *Covid* dataset.

Datasets	Covid→Mixed		
Transfer style	Pr	Re	F1
zero-shot	77.05	54.80	64.05
embedding-level	76.40	71.20	73.71
model-level	79.86	71.05	75.20
Datasets	Mixed→Covid		
Transfer style	Pr	Re	F1
zero-shot	36.23	45.85	40.41
embedding-level	67.19	68.85	68.01
model-level	66.44	71.47	68.86

Table 2.7: Cross-domain results for the BERT-BiLSTM-CRF-MTL model. Comparison between three transfer styles: zero-shot, embedding-level and model-level.

Model	Test	Pr	Re	F1
Model level transfer	MFT	79.86	71.05	75.20
	INV	67.78	52.29	59.03
Covid→Mixed	DIR	47.48	31.34	37.76
Model level transfer	MFT	66.44	71.47	68.86
	INV	57.64	59.35	58.48
Mixed→Covid	DIR	40.26	33.80	36.75

Table 2.8: Error analysis tests (MFT, INV and DIR) for the capability of the model-level transfer approach to generalize the concept of a location entity.

2.6.2 Cross-domain setting

Table 2.7 shows the results of the proposed BERT-BiLSTM-CRF-MTL model (which gives the best overall results in the in-domain setting) in the cross-domain setting. Specifically, we compare three transfer styles, *zero shot*, *embedding-level*, and *model-level*, when *Covid* is used as source and *Mixed* as target, and the other way around. As expected, the model-level transfer style gives the best results overall, while the zero shot style gives the worst results overall. Notably, in the case of the *Covid* to *Mixed* transfer, the model-level transfer improves the results of the in-domain model, from 74.67% to 75.20%. This is probably due to the diversity in the *Covid* dataset, which enables more accurate locations to be identified in the *Mixed* dataset (at a cost in recall). As opposed to that, the transfer from *Mixed* to *Covid* causes more specific locations to be identified, which improves the recall but negatively affects the precision (and the overall F1-measure).

No.	Test	Tweet text
1	MFT	NEW Hurricane Florence has made landfall near Wrightsville Beach, _{ctc→ctc} North Carolina _{sta→sta} at 7 15 AM EDT 1115 UTC with estimated maximum winds of 90 mph (150 km/h), and a minimum central pressure estimate of 958 mb (28.29”).
	DIR	NEW Hurricane Florence has made landfall near WI, _{sta→ctc} Panama _{ctc→sta} at 7 15 AM EDT 1115 UTC with estimated maximum winds of 90 mph (150 km/h), and a minimum central pressure estimate of 958 mb (28.29”).
2	MFT	On this Easter Sunday my thoughts are with Sri Lanka _{con→con} following the horrific attacks on worshippers there.
	INV	On this Easter Sunday my thoughts are with South Africa _{con→reg} following the horrific attacks on worshippers there.

Table 2.9: Examples of location predictions for different error analysis test settings (including MFT, INV and DIR). The examples are from the Mixed datasets and the predictions are made using model-level transfer from Covid to Mixed. The locations are highlighted with pink. The gold labels and model predictions for entities are highlighted in blue and shown as a subscript to the corresponding locations using the convention gold label→model prediction.

2.6.3 Error analysis

We performed error analysis of the model-level transfer approach from Table 2.7 (specifically, model-level transfer from *Covid* to *Mixed* and model-level transfer from *Mixed* to *Covid*). The analysis is based on the framework proposed by Ribeiro et al. (2020), where a model is tested for a capability using three tests: minimum functionality test (MFT), invariance test (INV) and directional expectation test (DIR). We performed the tests on the model’s capability to generalize the concept of a location entity. In our case, MFT is the model’s performance on the original *Mixed* or *Covid* test set, respectively. For INV, the location entities in the original test set were replaced with other randomly selected location entities of the same type from the test set. Finally, for DIR, the original location entities were replaced with randomly selected location entities of different types from the test set. The results of the analysis are shown in Table 2.8. The MFT score gives the performance on the original *Mixed* test set, and serves as a baseline for the other two tests. As can be seen, in both cases, the performance degrades when the locations are mixed up - tests INV and DIR as compared with the test MFT - suggesting that the model captures correlations between locations and their context. However, the F1 score for INV is better than the F1 score for DIR, which shows that the model expects a particular type of location in a given context.

Table 2.9 shows sample predictions for different tests (MFT, INV, DIR). In the first example, for the MFT test, the model makes a correct prediction for a tweet where a location entity of type *ctc* is followed by a location entity of type *sta*, which is the general convention for specifying a *city, state* location. However, for the DIR test, when the entities are replaced with others in reverse order of the type as compared to the original tweet (i.e., *sta, ctc* instead of *ctc, sta*), the model incorrectly, but not surprisingly, predicts *sta* as *ctc* and vice versa. In the second example, for the MFT test, the model correctly predicts Sri Lanka as a country (i.e., *con*). However, when *Sri Lanka* is replaced with *South Africa* in the case of the INV test, the model predicts it as *reg*. This is probably because Africa as a continent is a location of type *reg*, and also because cardinal directions are commonly associated with *reg* locations. Hence, without any external knowledge about *South Africa* as a country, *reg* is the next best prediction.

2.7 Conclusions

In this chapter, we introduced two new crisis tweet datasets manually tagged with specific fine-grained location types. These are the first manually annotated datasets for fine-grained location identification in crisis tweet texts, and can foster research in this area of great importance for crisis monitoring and response. The two datasets are different in nature, with one of them focused on mixed natural and man-made crisis events, which are generally localized to specific regions, and the second one focused on the worldwide Covid-19 pandemic. The different nature of the two datasets enables studies on location identification for localized and global events, as well as studies on the transferability of information between localized and global events.

In addition to introducing these datasets, we also designed a deep learning approach for fine-grained location identification, which combines BERT embeddings for input representation, BiLSTM as a context encoder, CRF as a tag decoder, and MTL to incorporate an auxiliary task. We studied this approach and showed its effectiveness and ability to transfer information between datasets.

Chapter 3

Multi-task learning to enable location mention identification in the early hours of a crisis event

3.1 Introduction

The huge amount of data that is posted on social media during crisis events could be used to build reliable and robust deep learning models for identifying information useful to crisis management and response teams. However, using social media data for a particular task, oftentimes, requires intensive manual effort in the form of annotation. The effort becomes even more arduous when we consider the noisy nature of social media content and the amount of labelled data required for a typical deep learning model.

The domain of crisis-related social media analysis, tweets in particular, is a well-researched field with labelled data available for various tasks ([Alam et al., 2018b](#); [Imran et al., 2016](#); [Middleton et al., 2014](#)). However, most of the available human-annotated datasets consists of thousands of instances, at best, which means that crisis datasets are relatively small compared to those available for tasks in other domains. Furthermore, for tasks that can support a new, emergent crisis situation, human-labelled data of large volume cannot be acquired for the reasons discussed above. In this

work, we explore ways in which we can harness the available small datasets from the domain to bolster performance for individual tasks of interest.

One popular approach in addressing the size-limitation of labelled data for a particular task is to leverage unlabelled data. In the field of Natural Language Processing (NLP), the recent advancements in transformer-based architectures, and associated pre-training with huge amounts of unlabelled data, has largely been successful in addressing this issue. Transformer-based architectures, currently, hold state-of-the-art results for many NLP tasks. However, the domain shift from the pre-training corpus to a downstream task’s domain is still a significant issue (Han and Eisenstein, 2019). Moreover, further pre-training with domain-specific unlabelled data is compute-intensive (Devlin et al., 2018) and it is not always feasible.

An alternative approach to address the limitation in terms of labeled data is to concurrently leverage smaller datasets available for different, but related tasks using a multi-task learning strategy (Caruana, 1997). In the multi-task setting, some layers can be shared across different tasks, while each task can also have one or more task-specific layers, and the entire model is trained in parallel for all the tasks. A multi-task model is designed with the intuition that the lower layers of the model learn abstract features common to related tasks, while the upper layers learn features specific to each individual task. This approach is especially useful in the domain of crisis-related social media analysis, given the lack of large datasets, while smaller datasets are available for different tasks.

In this work, our main focus is on the task of identifying fine-grained locations from tweet texts. Identification of location entities in tweets posted during crisis events is vital in extracting actionable situational awareness information. Furthermore, identifying the entities according to a hierarchy of location types can help in geographical location disambiguation and geo-coding. We use the English subset of the dataset published by Middleton et al. (2014), which has only a few thousands instances.

To address the issue of limited data size, we use a multi-task learning setting to augment the learning of fine-grained location identification with three other tasks in the domain of crisis-related tweets: key-phrase identification (Chowdhury et al., 2020), eyewitness-account classification (Zahra et al., 2020) and humanitarian categories classification (Alam et al., 2018b). We hypothesize that the similar nature of the tasks and the common abstract objective of identifying actionable

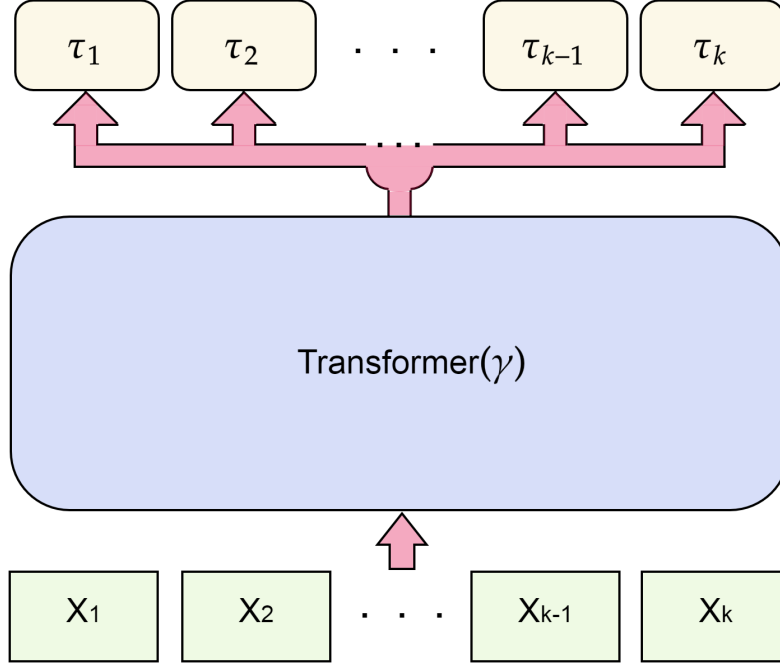


Figure 3.1: *Multi-task model overview: The architecture shows k task-specific modules (τ) attached on top of a shared Transformer (γ) module. The model is trained concurrently on all tasks with task-specific inputs (X).*

information in crisis-related tweets will result in a performance boost for the main task considered, i.e., identification of fine-grain locations in tweet texts.

3.2 Related work

Multi-task learning is a well-researched topic in the field of NLP, and deep learning, in general (Ruder, 2017). Caruana (1997) outlines one of the popular strategies for implementing multi-task learning: hard parameter sharing. In hard parameter sharing, a module shared by all tasks is followed by task-specific modules. On the other hand, in soft parameter sharing (Duong et al., 2015; Yang and Hospedales, 2016), each task has its own set of layers, but the parameters of the task-specific layers are constrained to be similar across tasks to enforce exchange of information among tasks. Hard parameter sharing is known to reduce over-fitting (Ruder, 2017), and is thus useful for tasks with small training sets. The usefulness of multi-task learning has been shown in a variety of applications, including image (Cheng et al., 2011; Zhang et al., 2014), voice (Rao

et al., 2018; Stoller et al., 2018) and text (Liu et al., 2019a; Wang et al., 2020) analysis. In the crisis domain, Wang et al. (2020) presented a multi-modal multi-task model using a single multi-modal dataset containing labels for different tasks. Chowdhury et al. (2020) used single-token keywords identification as an auxiliary task when predicting multi-token keyphrases. The work by Liu et al. (2019a) is the closest to our work, in that they used separate task-specific datasets in a multi-task learning setting. However, they performed experiments on general NLP benchmark datasets, such as GLUE (Wang et al., 2018).

3.3 Background and approaches

We use the hard-parameter sharing approach for multi-task learning, where task-specific modules (τ) are attached on top of a shared module (γ) as shown in Figure 3.1. We use the algorithm proposed by Liu et al. (2019a) to train multiple task modules, in parallel.

3.3.1 Multi-task learning

Multi-task learning can be formulated as follows: Given a set of k tasks and their corresponding data $D = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_k, Y_k)\}$, where $(X_i, Y_i) \in D$ is the training dataset for the i^{th} task, the goal of multi-task learning is to minimize the aggregate loss $L(\theta)$ on D given by:

$$L(\theta) = \sum_{i=1}^k \sum_{(x,y) \in (X_i, Y_i)} \ell^i(y, T^i(x, \theta)) \quad (3.1)$$

where ℓ^i is the loss function for the i^{th} task, $T^i(x, \theta) = \tau^i(\gamma(x, \theta_\gamma), \theta_i)$ is the output of the i^{th} task module τ^i (with parameter $\theta_i \subset \theta$) and γ is the shared module (with parameter $\theta_\gamma \subset \theta$).

3.3.2 Token classification tasks

A token classification task maps a sequence $x = \{x_1, \dots, x_n\}$, representing n input tokens, to a sequence $y = \{y_1, \dots, y_n\}$, representing n task-specific tags corresponding to the tokens in the sequence x . The mapping can be generically written as: $\tau^i(\theta_i) : \mathbf{x} \rightarrow \mathbf{y}$ (where θ_i are the task

parameters). The loss ℓ^i that we use for these tasks is the negative log likelihood. In our setting, the fine-grained location identification and the keyphrase identification tasks are token classification tasks.

3.3.3 Sequence classification tasks

A sequence classification task maps a sequence of n tokens $x = \{x_1, \dots, x_n\}$ to a class $y \in C$, where C is the task-specific set of classes. The mapping can be generically written as $\tau^j(\theta_j) : \mathbf{x} \rightarrow \mathbf{y}$ (where θ_j are the task parameters). We use the standard cross-entropy or the binary cross-entropy loss ℓ^j for sequence classification tasks, depending upon the number of classes. In our setting, the eyewitness-account classification and the humanitarian categories classification tasks are seen as sequence classification tasks.

3.3.4 Shared module

For the shared module, we experiment with three base variants (hidden layer size of 768) of transformer-based language models: BERT (Devlin et al., 2018), Albert (Lan et al., 2019) and RoBERTa (Liu et al., 2019c). BERT is one of the first and most popular transformer-based models, and has been used widely in various applications. We compare its performance with that of RoBERTa, which was trained on a larger dataset than BERT and it is considered to be more robust (Liu et al., 2019c). We also compare BERT and RoBERTa with Albert, a significantly smaller model, to understand the effect that the size of the shared module has on the the performance of the task at hand.

3.4 Experimental setup

3.4.1 Datasets

Fine-grained location identification

We use the dataset published by [Middleton et al. \(2014\)](#), which has two sets of tweets posted during Hurricane Sandy 2012 and Christchurch Earthquake 2012, respectively. Based on the place of the event, in what follows, the set of tweets posted during Hurricane Sandy 2012 will be referred to as *NY*, while those posted during Christchurch Earthquake 2012 will be referred to as *NZ*. The NY and NZ sets contain 1907 and 1762 unique tweets, respectively. Both the NY and NZ tweets have human-labelled location entities corresponding to three categories: *administrative location*, *building* and *transportation*. We randomly select 1000 tweets from each set for training, 500 tweets for test, and the remaining tweets for development (dev) splits.

Keyphrase identification

We use the dataset published by [Chowdhury et al. \(2020\)](#) for keyphrase identification. The dataset contains tweets from various crisis events. The tokens in each tweet are labelled as keyphrase or not, using the script provided by [Chowdhury et al. \(2020\)](#). We use a random sample of 1000 tweets for training to keep the dataset size balanced across multiple tasks.

Eyewitness-account classification

The dataset published by [Zahra et al. \(2020\)](#) contains tweets from flood, hurricane and earthquake. The tweets are labelled using one of five eyewitness classes: *direct-eyewitness*, *indirect-eyewitness*, *vulnerable direct-eyewitness*, *non-eyewitness*, and *don't know*. As for the other tasks, we use a class-balanced sample of 1000 tweets from the dataset for training.

Humanitarian categories classification

We use the dataset published by [Alam et al. \(2018b\)](#) for humanitarian categories classification. The dataset contains tweets from various crisis events, labelled using the following humanitarian classes:

Infrastructure and utility damage, Vehicle damage, Rescue, volunteering, or donation effort, Injured or dead people, Affected individuals, Missing or found people, Other relevant information, and Not relevant or can't judge. We use a class-balanced sample of 1000 tweets from the dataset for training.

3.4.2 Baseline and metrics

We use a single task-setting, where only one task-module is attached over the base module, as the baseline to which we compare the performance of the multi-task learning setting. We use precision (Pr), recall (Re) and F1 scores as metrics to compare the performance of the models.

3.4.3 Experiments and implementation details

Hyperparameter	Values
Learning rate	1e-5
Learning rate decay	Linear
Batch size	{4, 8, 16}
Classifier Dropout	0.1
Optimizer	AdamW
Adam epsilon	1e-8
Maximum Epochs	50

Table 3.1: *Hyperparameter configuration used in the experiments.*

We perform the experiments on three levels of task-sharing: a **Single-task model (ST)** (trained on *NY* and *NZ*, respectively, and tuned/tested on the corresponding dev/test sets), a **Location-only Multi-task model (LMT)** (trained with *NY* and *NZ* as two different location identification tasks in a multi-task setting, and tuned/tested on the corresponding dev/test sets), and **Multi-task model (MT)** (trained with all tasks in a multi-task setting, and tuned/tested on the corresponding dev/test sets for the location identification tasks). In addition, we also combine the two location datasets *NY* and *NZ* (*Combined*) and train a single-task model (*ST*), and a multi-task model (*MT*), using the other three tasks to understand the benefit of only using the related tasks for general location identification (as opposed to disaster-specific identification). The hyperparameter configuration used

Task	γ	Dev-F1	Test		
			Pr	Re	F1
Dataset: NY					
ST	BERT	75.40	79.18	79.41	79.30
	RoBERTa	66.31	70.82	73.53	72.15
	Albert	72.40	78.72	76.18	77.43
LMT	BERT	75.49	76.09	82.35	79.10
	RoBERTa	69.69	81.72	69.71	75.24
	Albert	73.29	76.68	77.35	77.01
MT	BERT	73.95	78.90	80.29	79.59
	RoBERTa	70.12	76.95	69.71	73.15
	Albert	77.43	84.00	74.12	78.75
Dataset: NZ					
ST	BERT	67.68	58.55	66.86	62.43
	RoBERTa	71.25	66.67	57.99	62.03
	Albert	69.52	61.58	69.23	65.18
LMT	Bert	65.99	56.72	67.46	61.62
	RoBERTa	69.01	66.67	63.91	65.26
	Albert	70.83	70.35	71.60	70.97
MT	BERT	72.32	69.05	68.64	68.84
	RoBERTa	70.37	72.79	58.58	64.92
	Albert	75.14	67.47	66.27	66.87
Dataset: Combined					
ST	BERT	75.95	73.14	85.07	78.66
	RoBERTa	69.90	72.59	68.17	70.31
	Albert	68.92	71.72	68.76	70.21
MT	BERT	76.33	76.01	80.94	78.40
	RoBERTa	69.01	71.98	70.14	71.04
	Albert	78.65	83.54	77.80	80.57

Table 3.2: Results for Fine-grained location identification task in three settings: Single-task (ST), Location-only Multi-task (LMT) and all-multi-task (MT) with three variants of transformers (γ) as shared module. The results are grouped by dataset: NY, NZ and Combined.

in the experiments are shown in the Table 3.1. All experiments are run on a 4-GPU system of NVIDIA Tesla V100 GPUs.

3.5 Results and discussion

Table 3.2 shows the results for the location identification task grouped by dataset (NY, NZ and Combined) in three different task-sharing settings, with three variants of the transformer-based shared module, as described in Section 3.4.3. The table shows precision, recall and F1 scores for

Transformer	Average performance improvement (%)	
	ST → LMT	ST → MT
BERT	-0.78	3.43
RoBERTa	4.75	2.36
Albert	4.17	6.35

Table 3.3: Average (macro) performance improvement over three datasets for the three transformer variants, when comparing single-task model (ST) to location-only multi-task (LMT) and all-multi-task (MT) model.

the test set, along with F1 score on the corresponding development set (Dev-F1) used to identify the best model.

As can be seen in Table 3.2, the multi-task models (MT) consistently outperform the corresponding single-task models (ST) in terms of test F1 score, with the exception of the *Combined* set, when BERT is used as the base module. It is interesting to see that the models, however, have very little or no gain when only using identical tasks (LMT). On the other hand, we see the highest overall performance improvement of 14.76% for MT as compared to ST, on the *Combined* set, with Albert as shared module, and all the non-location tasks.

Table 3.3 shows the macro average of percentage improvement in F1-score from single-task models to the two variants of multi-task models for the three transformer-based modules. In general, we see significant performance improvement from the single-task to the multi-task model. From the table, Albert has the best overall performance improvement for the task of location identification. This could be because Albert has significantly less parameters compared to the other two variants, and given the small size of the datasets, that results in better generalization and shared learning. The overall lower performance gain from ST to LMT could be due to the fact that the two tasks are very similar and thus, they do not help model’s generalization capability. On the other hand, when using more dissimilar tasks, the model benefits from variations in the data and is able to generalize better; this results in performance improvement across all variants.

3.6 Conclusions

In this paper, we studied the effect of using other closely related tasks in a multi-task setting for the task of fine-grained location identification in the domain of crisis-related tweets. Our results show that using multi-task learning can improve the performance significantly as compared to single-task learning. The approach could be specially useful when analysing an emergent crisis, where annotating large amount of data is not feasible and the entire analysis process is time-critical. Moreover, our results also show that when a small amount of data is available, the model may benefit more from using other related tasks rather than using additional data for the same type of task (i.e., two location identification tasks).

In addition to studying the effect of multi-task learning, we also studied the effect of different transformer variants on the performance of the model for the main task. With different variants available, it is necessary to find the best-fitting one for the task at hand. The results show that the choice of the shared module has significant effect on the performance.

Chapter 4

Contrastive learning for multimodal classification of crisis related tweets

4.1 Introduction

Identifying informative content from social media during a crisis event is an important step towards making use of the huge influx of data in building an actionable knowledge base. The task of finding informative content is particularly challenging in social media because social media posts, tweets in particular, are inherently noisy in nature and rife with irrelevant or inaccurate information ([Hunt et al., 2020, 2022](#); [Wang et al., 2019](#)). Besides the trustworthiness of the content, another challenge with tweets, and digital social media content in general, is the multimodality of the data. A tweet is usually a short text accompanied with other media like images and videos. However, most prior works have primarily focused on tweets as a text-only problem ([Glandt et al., 2021](#); [Stowe et al., 2016](#); [To et al., 2017](#)). Given the 280 character limit on the text of a tweet, the use of associated image as an additional modality of data is a natural extension and has shown improved performance on a variety of tweet-related tasks beyond the domain of crisis informatics ([Adjali et al., 2020](#); [Wang et al., 2022](#)).

One of the main challenges in building a multimodal system is aligning the representation of various modalities into a common latent space. One typical approach is to learn the representation

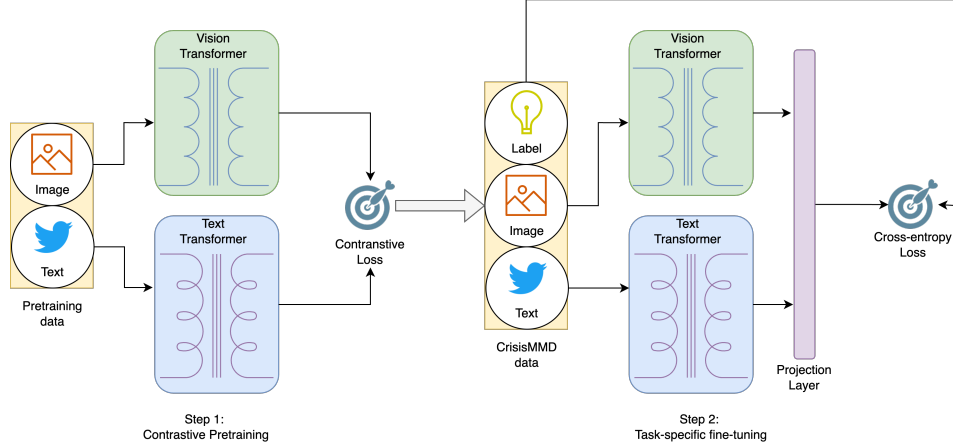


Figure 4.1: Overview of the two-step training process with contrastive pre-training and task-specific fine-tuning on CrisisMMD.

for each modality independently and then use different fusion strategies as part of downstream task fine-tuning to align the representations (Gautam et al., 2019; Madichetty et al., 2020; Xukun and Caragea, 2020). There have been significant advancements in self-supervised representation learning of individual modalities with models like CNN (Simonyan and Zisserman, 2014) and Transformers (Vaswani et al., 2017). However, self-supervised joint representation learning for multiple modalities is still in its early stage by comparison.

Contrastive learning is a learning paradigm where the model learns representations by contrasting pairs of similar or dissimilar data samples. The aim of contrastive learning is to learn representations where similar concepts are grouped closer in the latent vector space, while dissimilar concepts are moved further away. Radford et al. (2021) applied this technique to text-image pairs to learn a joint representation for the two modalities. One key assumption made in training this model is that the text directly explains the content of the image. This is generally true for the caption-image pairs used in training the model. However, the relationships between text and images in tweets are more nuanced and complex (Vempala and Preoȃuc-Pietro, 2019). In the domain of crisis related tweets, Sosea et al. (2021) created a dataset of text-image pairs with their relationships. Biamby et al. (2021) used pre-trained models provided by Radford et al. (2021) to identify misinformation in tweets and found that the relationship type between image-text pairs affects the performance of the model. In this work, we show that additional domain-specific contrastive pre-training helps

to learn the nuanced text-image relationship in disaster related tweets and consequently improves performance on the downstream tasks. In particular, we use the task of tweet informativeness classification and humanitarian categories classification from *CrisisMMD* (Alam et al., 2018b), a publicly available multimodal dataset of tweets posted during various disasters, as the downstream tasks. The overview of the two-step training process with domain-specific contrastive pre-training and task-specific fine-tuning is shown in Figure 4.1.

4.2 Related work

Multimodal deep learning has been used for a variety of tasks and modalities, including audio-visual multimodal learning for speech classification (Ngiam et al., 2011) and emotion recognition (Chen and Jin, 2015; Wöllmer et al., 2010), image-text for image captioning (Bernardi et al., 2016), question answering (Antol et al., 2015) and text-image matching (Socher et al., 2013), etc. In the context of tweets, multimodal learning is generally applied to image-text pairs. Specifically, it has been used in tasks such as sentiment analysis (Kumar and Garg, 2019; Yu and Jiang, 2019), mental health analysis (Gui et al., 2019; Kang et al., 2016; Safa et al., 2022), event detection (Cai et al., 2015; Yilmaz and Hero, 2018), etc.

For crisis related tweets, *CrisisMMD* (Alam et al., 2018b) is one of the most popular multimodal datasets. The CrisisMMD dataset contains tweets posted during seven different natural and human-induced crisis events and has two multimodal tasks: *tweet informative* and *humanitarian categories*. More details about the CrisisMMD tasks are provided in the dataset section. There has been a variety of multimodal learning approaches applied to this dataset. The most popular approaches can be seen as prediction over fusion of embeddings with separate encoders for each modality. Almost all the approaches use a variation of pre-trained image encoder based on convolutional neural network (CNN) and different text encoders, including word-embeddings (Madichetty et al., 2020; Nalluru et al., 2019; Zou et al., 2021), recurrent neural networks (RNN) (Gautam et al., 2019; Xukun and Caragea, 2020), CNNs (Chen et al., 2020b; Ofli et al., 2020) and transformers (Abavisani et al., 2020; Madichetty et al., 2021; Vaswani et al., 2017).

Prior works employ a variety of fusion techniques for merging the individual embeddings of

different modalities. [Madichetty et al. \(2020\)](#) and [Madichetty et al. \(2021\)](#) use additive fusion where the class probabilities predicted for each modality are added to obtain the multimodal probability prediction. Another common approach is to concatenate the embeddings of the constituent modalities before feeding them to a classifier unit ([Chen et al., 2020b](#); [Gautam et al., 2019](#); [Nalluru et al., 2019](#); [Ofli et al., 2020](#); [Xukun and Caragea, 2020](#); [Zou et al., 2021](#)). [Abavisani et al. \(2020\)](#) use a cross-attention module over individual embeddings as the classifier unit.

For the *CrisisMMD* tasks, we observed a lack of use of the standard benchmark splits provided by [Ofli et al. \(2020\)](#) amongst prior works. We observed that some works report results for different crisis events separately. For the ones that report the results separately, some use all seven events ([Gautam et al., 2019](#); [Madichetty et al., 2021, 2020](#); [Xukun and Caragea, 2020](#)) while others use a subset of the events ([Chen et al., 2020b](#); [Nalluru et al., 2019](#)). We also observed a discrepancy in terms of how ground truth labels are decided for a text-image pair where the labels for the text and image modalities may be different. Some works use the text labels only ([Madichetty et al., 2020](#)), while others use the “informative” label when either the text or the image are labelled so ([Chen et al., 2020b](#); [Madichetty et al., 2021](#)). However, we found that the most common approach is to use the pairs where the labels for image and text match ([Abavisani et al., 2020](#); [Gautam et al., 2019](#); [Nalluru et al., 2019](#); [Xukun and Caragea, 2020](#); [Zou et al., 2021](#)).

Self-supervised representation learning for model pre-training has been one of the most influential developments in deep learning in recent years ([Deng et al., 2009](#); [Vaswani et al., 2017](#)). Such pre-trained models can be subsequently fine-tuned on downstream tasks with relatively less data to achieve good performances. As can be seen in prior works, multimodal representation learning, however, has mostly been perceived as part of task-specific training. Similar to our work, [Shome and Kar \(2021\)](#) use multimodal contrastive pretraining followed by task-specific training to identify offensive tweets. However, they use supervised contrastive learning where text-image pairs belonging to the same class are contrasted against those belonging to a different class. This approach does not scale well as it requires annotated data and is unable to make use of unlabelled data. [Radford et al. \(2021\)](#) proposed a self-supervised joint representation learning approach for text-image pairs with multimodal contrastive learning called Contrastive Language-Image Pre-training (CLIP). CLIP does not require label for the text-image pairs to create negative samples. CLIP has shown

promising results in variety of tasks. Most notably, CLIP is an integral part of some of the popular text-to-image models like DALL-E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022) and Stable Diffusion (Rombach et al., 2021). Bielawski et al. (2022) used a pre-trained CLIP model for transfer learning for the task of sentiment analysis in tweets. They conclude that CLIP outperforms the baseline approaches in *human-centric* tasks as the model is pre-trained on labels generated by humans for other humans as opposed to crowd-sourced annotations where the labels are generated by human to specifically train machines. They define a task as *human-centric* “whenever a concept refers to cultural, social, aesthetic and/or affective components of the world” (Bielawski et al., 2022). Based on this definition, both informativeness of a tweet and the categorization of humanitarian classes in *CrisisMMD* can be thought of as *human-centric* tasks.

4.3 Background and problem definition

We use the CLIP approach proposed by Radford et al. (2021) for self-supervised multimodal representation learning, i.e., multimodal model pre-training. For each batch of data containing N text-image pairs, CLIP creates $N^2 - N$ additional pairs by cross-mapping all the images and the corresponding texts in the batch - these additional pair can be thought of as negative samples. The objective is to maximize the similarity score for the true pairs.

We use the OpenAI vision transformer (Radford et al., 2021) variant of the CLIP pre-trained model. We perform pre-training using the contrastive loss objective and fine-tune the model on the informative and humanitarian tasks from *CrisisMMD*. The details of the pretraining dataset and fine-tuning dataset are described in Section 4.4. For fine-tuning the model, we concatenate the encoder outputs of the vision and text transformers before feeding it of a dense layer for classification.

Both informative and humanitarian tasks from *CrisisMMD* are classification tasks. The dataset contains individual labels for each modality of data. Formally, given a dataset with text-image pairs (t_x, i_x) , the task is to find a mapping function $f_\theta : (t_x, i_x) \rightarrow (t_y, i_y)$ from the input tuple to output tuple (t_y, i_y) where t_y is the label corresponding to t_x and i_y is the label corresponding to i_x . However, following up on prior works (Abavisani et al., 2020; Gautam et al., 2019; Ofli et al., 2020; Zou et al., 2021), we simplify the problem by considering a single label for the pair. Hence, the

	Train (70%)		Dev (15%)		Test (15%)		Total	
	Text	Image	Text	Image	Text	Image	Text	Image
Informative								
Informative	5,546	6,345	1,056	1,056	1,030	1,030	7,632	8,431
Not-informative	2,747	3,256	517	517	504	504	3,768	4,277
Total	8,293	9,601	1,573	1,573	1,534	1,534	11,400	12,708
Humanitarian								
Affected individuals	70	71	9	9	9	9	88	89
Rescue volunteering or donation effort	762	912	149	149	126	126	1,037	1,187
Infrastructure and utility damage	496	612	80	80	81	81	657	773
Other relevant information	1,192	1,279	239	239	235	235	1,666	1,753
Not-humanitarian	2,743	3,252	521	521	504	504	3,768	4,277
Total	5,263	6,126	998	998	955	955	7,216	8,079

Table 4.1: Dataset sizes for informative and humanitarian tasks from *CrisisMMD* as provided in the benchmark splits by [Ofli et al. \(2020\)](#).

updated task is a mapping function $f_\theta : (t_x, i_x) \rightarrow y$ where y is the unique label corresponding to the input pair.

4.4 Datasets

4.4.1 CrisisMMD

We use *CrisisMMD*, a publicly available multimodal dataset of tweets posted during various crisis events. The dataset contains tweets posted during seven events of natural disasters like hurricanes, floods, earthquakes and wildfires. We describe the dataset and its use in our experiments in the remainder of this section.

The *Informativeness* task in *CrisisMMD* dataset contains two labels: “*Informative*” and “*Not informative*” for each text and image that are independently labelled.

For the *Humanitarian* task in *CrisisMMD*, each text and image are labelled independently as one of eight humanitarian categories: “*infrastructure and utility damage*”, “*not humanitarian*”, “*other relevant information*”, “*rescue volunteering or donation effort*”, “*vehicle damage*”, “*affected individuals*”, “*injured or dead people*” and “*missing or found people*”.

Pretraining set \ Source	CrisisMMD		DisRel		Total	
	Train (80%)	Dev (20%)	Train (80%)	Dev (20%)	Train	Dev
CrisisMMD Only	5,754	2,465	-	-	5,754	2,465
CrisisMMD+DisRel	5,774	2,445	3,473	1,518	9,247	3,963

Table 4.2: Distribution of pretraining dataset for two settings: *CrisisMMD only* and *CrisisMMD+DisRel*

We use the benchmark splits provided by [Offi et al. \(2020\)](#), for which the distribution is shown in Table 4.1. Since the dataset was annotated for the two modalities separately and independently and there can be differing labels for a single pair, the splits only contain tweets with matching labels for image and text. Furthermore, for the *Humanitarian* task, semantically similar classes are merged together. In particular, “*injured or dead people*” and “*missing or found people*” are merged into “*affected individuals*”, and “*vehicle damage*” is merged into “*infrastructure and utility damage*”.

4.4.2 Pre-training dataset

For pre-training the model on tweets with text-image pairs, we collected text-image pairs from two sources. Firstly, we collected text-image pairs from the train splits from Table 4.1. Since the splits for the two tasks are created independently, there is some overlap between the training and development/testing sets across tasks. Hence, we use text-image pairs that are present only in the training set of either task. In addition to the tweets from *CrisisMMD*, we also use the text-image pairs from *DisRel* ([Sosea et al., 2021](#)), a multimodal dataset of tweets posted during various crisis events. Each text-image pair in *DisRel* is labelled as either “*Related*”, “*Complementary*” or “*Unrelated*”. We only use the pairs that are labelled as either “*Related*” or “*Complementary*”. We create a pre-training dataset with only *CrisisMMD* data and another one where we merge the data from the two sources. In both the settings, we create splits of 80:20 ratio for training and validation. The objective of the domain-specific pre-training is to learn the nuanced relationship between text-image pairs in tweets. We use the development set from *DisRel* dataset to study the effect of this pre-training in learning the relationships. Furthermore, with the pre-training dataset containing pairs from both *CrisisMMD* and *DisRel*, we want to study the effect of additional pre-training data with known relationships.

The distribution of image-text pairs for each source and setting is shown in Table 4.2.

4.5 Experiments

We pre-train the model with contrastive learning on the two pre-training datasets and use the models with the smallest validation loss for fine-tuning on the downstream tasks.

We study the possibility of few-shot learning with the pre-trained models by fine-tuning on the downstream tasks using a fraction of the training data. Specifically, we train the model by taking random sample with replacement of multiple of 10% of data. The goal of this experiment is to study the effect of pre-training in achieving comparable performance with respect to the baseline models with less training data. This can be useful, especially, in the domain of crisis informatics because it is not always feasible to gather carefully annotated data for an emergent crisis.

We compare the results with various other multimodal models as explained in the following section.

4.5.1 Baseline and proposed models

CLIP

We use a vision transformer variant from OpenAI (Radford et al., 2021), specifically *clip-vit-base-patch32*. The encoder is pre-trained using CLIP algorithm on caption-image pairs. We use this model as baseline for CLIP based models without additional domain-specific pre-training.

CLIP pre-trained with *CrisisMMD only*

We use the CLIP model above and pre-train it using the *CrisisMMD Only* pre-training set.

CLIP pre-trained with *CrisisMMD+DisRel*

We use the CLIP model above and pretrain it using the *CrisisMMD+DisRel* pre-training set.

VGG-16+CNN (Ofli et al., 2020)

The model uses VGG-16 as image encoder and CNN over Word2vec as text encoder. The model uses concatenation of the separate embeddings as joint representation that is fed it to a dense layer for classification. For the *CrisisMMD* task, this is a standard baseline model.

DenseNet+BERT (Abavisani et al., 2020)

The model uses DenseNet and BERT as image and text encoders, respectively. The authors employ cross-attention between the text and image embeddings with stochastic transition between the embeddings of the two modalities to ensure robustness. The paper reported the best results in our literature review for an arbitrary split. We use an unofficial implementation ¹ of the paper to train the model on the benchmark dataset splits and report the results.

4.5.2 Performance metrics

We use the standard accuracy, precision, recall and F1 metrics for evaluation of the models on the downstream tasks. We use support-weighted average of the scores because of the high class-imbalance in the dataset. For the comparison of the contrastively pre-trained models, we use the distribution of cosine similarity scores between the text and image embeddings for different relationship types.

4.6 Results and discussions

4.6.1 CLIP variants (pre-trained models)

Table 4.3 shows the comparison of the three CLIP variants used in our experiments in terms of the alignment of multimodal representation for different relationship types of text-image pairs. We use the *DisRel* development subset from Table 4.2 to compare the contrastively learned embeddings using cosine similarity between the text and image embeddings. The table shows the distribution of

¹<https://github.com/PaulCCCCCH/Multimodal-Categorization-of-Crisis-Events-in-Social-Media>

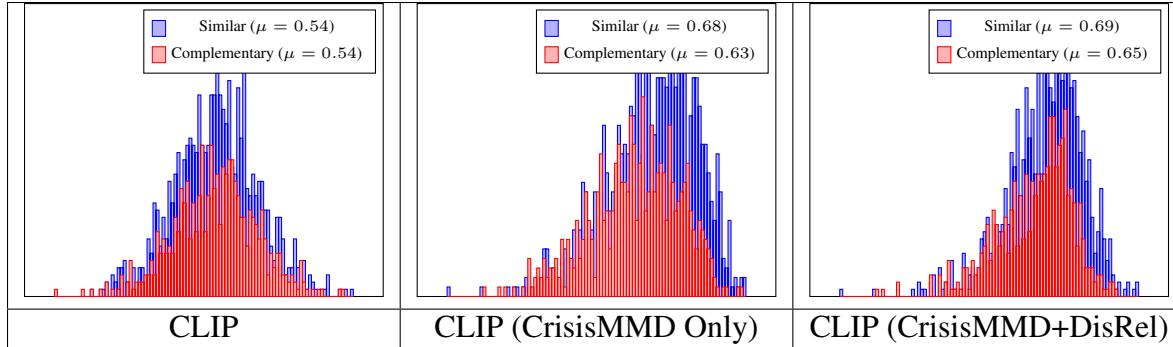


Table 4.3: Distribution of normalized cosine similarity for “Similar” and “Complementary” text-image pairs from DisRel pre-training development set for the three pre-trained variants of the CLIP model with mean score shown as μ .

the similarity scores for the “Similar” and “Complementary” text-image pairs from the set. As can be seen in the table, the rightward skew for the domain-specific pre-trained variants, CLIP (CrisisMMD Only) and CLIP (CrisisMMD+DisRel), suggests that the multimodal embeddings for the text-image pairs in the dataset are aligned better with the additional pre-training. Moreover, for the vanilla CLIP variant, the distributions for both *Similar* and *Complementary* pairs are similar with the same mean (μ) of 0.54. However, for the domain-specific pre-trained variants, the *Similar* pairs are scored higher on average than *Complementary* pairs, suggesting that the model learns the difference between these pair types. Furthermore, the scores for the CLIP (CrisisMMD+DisRel) is higher on average than the scores for CLIP(CrisisMMD Only), suggesting that the additional pre-training data of known relation types further improves the multimodal representation alignment.

4.6.2 CrisisMMD tasks

Table 4.4 and Table 4.5 show the result of the proposed model and the baseline models for the *Informative* and *Humanitarian* tasks, respectively. For both tasks, we can see that the CLIP model exceeds the performance of VGG-16+CNN baseline. For the *Informative* task, the CLIP model pre-trained with *CrisisMMD Only* performs the best with the weighted F1 score of 90.15. For the *Humanitarian* task, the CLIP model pre-trained with *CrisisMMD+DisRel* performs the best with the weighted F1 of 87.88, and the CLIP model pre-trained with *CrisisMMD Only* has second best score of 87.54.

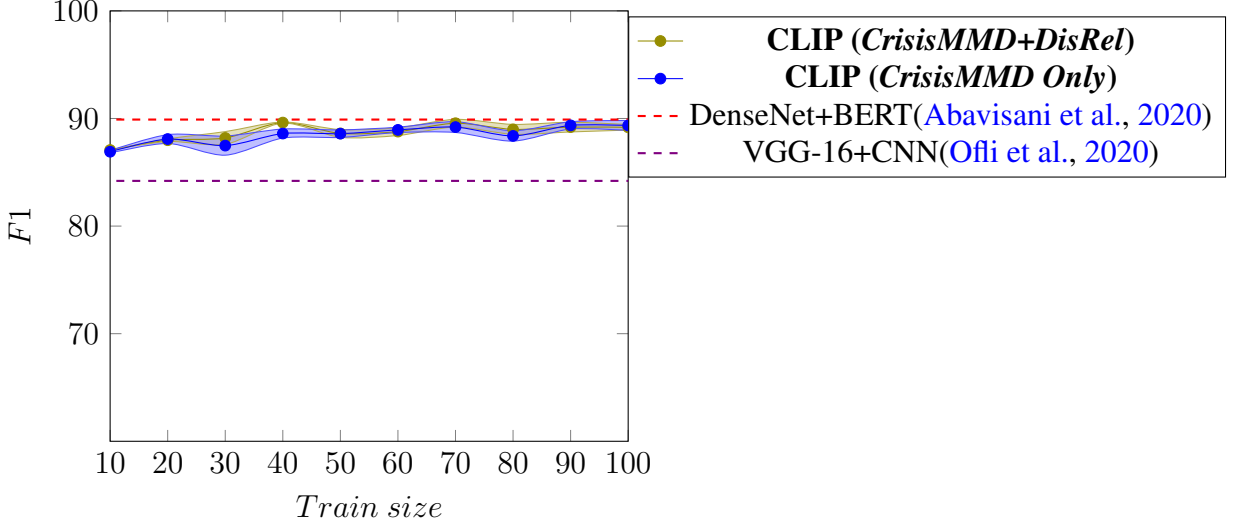


Figure 4.2: *F1 score on Informative task at varying training size for our proposed model. The baseline scores (at 100% training size) are shown as dashed line.*

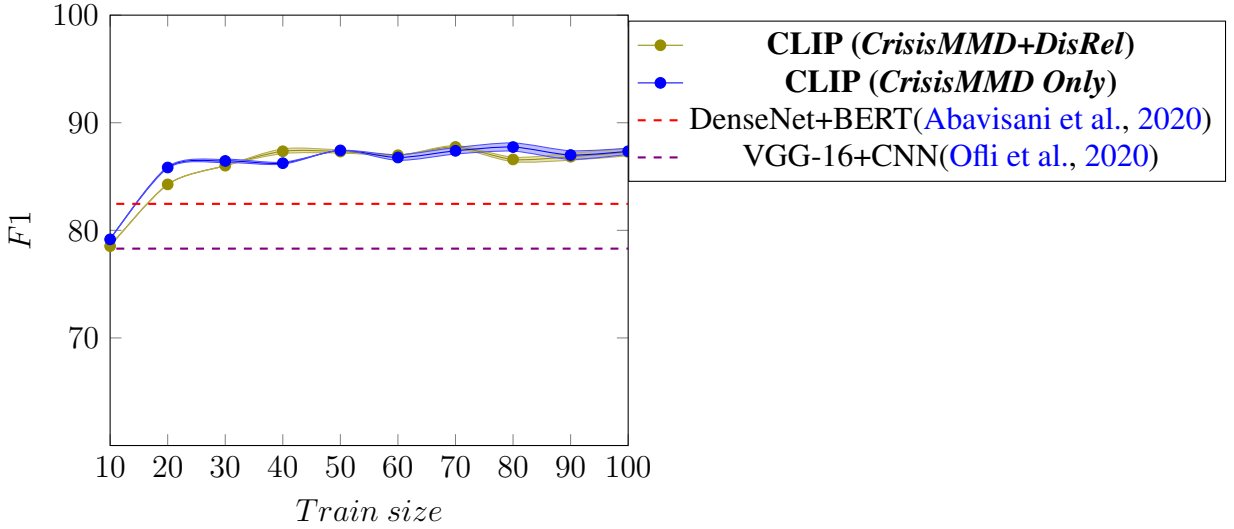


Figure 4.3: *F1 score on Humanitarian task at varying training size for our proposed model. The baseline scores (at 100% training size) are shown as dashed line.*

We can also observe that the DenseNet+BERT model outperforms the baseline VGG-16+CNN but at least one pre-trained variant performs better than the model for either task. It is also interesting that the CLIP model without any additional pre-training is able to achieve competitive score. However, domain specific pre-training helps improve the performance. This suggests that the model is transferable across domain but a small domain-specific pre-training helps. Figure 4.2 and Figure 4.3 show the result of few-shot learning for the two tasks for five trials. The thickness

Model	Accuracy	Precision	Recall	F1
VGG-16 + CNN (Ofli et al., 2020)	84.40	84.10	84.00	84.20
DenseNet + BERT (Abavisani et al., 2020)	89.90	89.92	89.90	89.90
CLIP	89.31	89.22	89.31	89.17
CLIP pre-trained with <i>CrisisMMD only</i>	90.29	90.23	90.29	90.15
CLIP pre-trained with <i>CrisisMMD + DisRel</i>	89.64	89.64	89.64	89.42

Table 4.4: Results of various baselines models and our CLIP based model on Informative task.

Model	Accuracy	Precision	Recall	F1
VGG-16 + CNN (Ofli et al., 2020)	78.40	78.50	78.00	78.30
DenseNet + BERT (Abavisani et al., 2020)	82.72	82.50	82.72	82.46
CLIP	87.76	86.86	87.76	87.23
CLIP pre-trained with <i>CrisisMMD only</i>	87.96	87.13	87.96	87.54
CLIP pre-trained with <i>CrisisMMD + DisRel</i>	88.27	87.57	88.27	87.88

Table 4.5: Results of various baselines models and our CLIP based model on Humanitarian task.

of the line plot at each node represents the standard deviation for the corresponding dataset size. Interestingly, for both tasks, and both pre-training settings, CLIP based models achieve better score than the baseline VGG-16+CNN with only 10% of the data. For the *Humanitarian* task, the CLIP based models achieve better score than the DenseNet+BERT model with only 20% of the total data. However, for the *Informative* task, the proposed model requires all the data to match the performance of DenseNet+BERT.

Among the two tasks, *Humanitarian* task is a relatively more complex task (Ofli et al., 2020) with more significant class imbalance. Interestingly, the proposed model’s performance in this task is also significantly better as compared to the baseline models, whereas for the simpler *Informative* task, the performance is relatively similar. This could be because of the simple nature of the task, basic multimodal representation is sufficient to achieve good results for the *Informative* task. However, for a complex task, contrastively learned joint multimodal representations help improve the performance.

4.7 Conclusions

Joint multimodal representation learning of text-image pairs has been successfully used in a variety of tasks recently. In this work, we used contrastive language-image pre-training to jointly learn representation for text-image pairs from tweets posted during various crisis events. Firstly, we showed that additional domain-specific contrastive pre-training helps learn different text-image relationships in tweets and achieves better multimodal representation alignment. For the two tasks, the proposed CLIP variants achieve the best scores. With the few-shot learning experiment, we observed that the CLIP trained models require relatively less fine-tuning data to match the results of the baseline models. Hence, contrastively learned joint multimodal representations can be used to build models with only a small amount of labelled data.

Chapter 5

Using content analysis and machine learning to identify COVID-19 information relevant to low-income households on social media

5.1 Introduction

The impacts of the COVID-19 pandemic have been especially severe for vulnerable populations, such as low-income households (LIHs). Due to staying-at-home orders and the speed of COVID-19 spread, low-income communities have suffered from the problems of unemployment, disease and illness, online education attainment, energy and internet accessibility and affordability, psychological stress, and overall well-being ([Beard, 2020](#); [Benfer et al., 2020](#); [Dahab et al., 2020](#); [Deaton, 2021](#); [Galasso, 2020](#); [Harris and Lewis, 2020](#); [Parker et al., 2020](#)). [Kreuter et al. \(2020\)](#) analyzed more than 3.5 million requests sent during COVID-19 to 2-1-1, a helpline that provides information about social services that address basic human needs (such as clothing, food, housing, utilities), mental health, children, youth and family support, among others. The results of the analysis strongly supported the claim that the social and economic impacts of COVID-19 are dramatic for low-income populations, while also showing that the needs observed vary across different communities. Furthermore, the authors emphasized that the “most social needs are not systematically tracked

by government, let alone in real time” (except for unemployment claims). Therefore, there is a compelling need for tools that can automatically and effectively collect information about LIH in real time.

One potential source of real-time information about LIH can be found in social media. Given the social distancing and physical isolation imposed by COVID-19, people have turned to social networking platforms to regain a sense of unity, communicate with family and friends, express opinions or complains, and even ask for emotional or financial help. [Tsao et al. \(2021\)](#) performed a review of the recent literature focused on the use and role of social media during COVID-19, and identified six public health themes that are prevalent in a large number of articles reviewed, including surveillance of public attitudes, identification of infodemics, assessing mental health, detection or predicting COVID-19 cases, analyzing government response to pandemic and evaluating the quality of information. The review also highlighted a lack of tools that could enable automatic real-time extraction of useful information from social media, and in particular, a lack of automated applications of machine learning approaches to COVID-19 social media data.

Most of the existing works on applying machine learning to COVID-19 social media data have focused on topic analysis, and sentiment or emotion analysis. For example, [Wicke and Bolognesi \(2021\)](#) and [Zhu et al. \(2020\)](#) used the Latent Dirichlet Allocation (LDA) ([Blei et al., 2003](#)) approach to identify prevalent COVID-19 topics and studied their evolution in time and/or space. [Nemes and Kiss \(2021\)](#) used recurrent neural networks (RNN) to detect the sentiment of COVID-19 tweets. Similarly, [Imran et al. \(2020\)](#) used long short-term memory (LSTM) networks to perform sentiment and emotion analysis, while [Das and Kolya \(2021\)](#) used convolutional neural networks (CNN) to predict the sentiment.

Generally, for tasks such as sentiment and emotion analysis, existing labeled datasets can be used to train supervised models that perform well in a variety of domains. However, for classification tasks that are more specific to COVID-19, training data need to be specifically labeled (or pseudo-labeled, e.g., using hashtags). While large datasets consisting of unlabeled COVID-19 social media have been assembled and made publicly available in the last year, there is still a lack of labeled datasets that can facilitate the extraction of specific COVID-19 information, and in particular information relevant to LIHs. This is due in part to the fact that there is a lack of understanding

regarding the information relevant to LIHs available on social media, and also to the fact that labeling data is laborious and expensive.

To address these limitations, we aim to utilize Twitter data contributed by affected individuals to gain insights into the public opinions toward the COVID-19 pandemic from the LIH perspective, by investigating key evidence that supports issues relevant to social order, crisis communication, and non-physical health impacts. More specifically, we approach the topic of COVID-19 from a conceptual framework of crisis communication ([Houston et al., 2015](#)) and theory of social order ([Elster, 1989](#)), with the goal of understanding people's responses to the social-psychological, behavioral and economic impacts of COVID-19. Subsequently, equipped with such understanding, we use the information categories that are well-represented on the Twitter social media platform to train machine learning classifiers that can automatically detect information relevant to LIHs in tweet streams. Thus, our contribution is twofold, as we integrate social science's content analysis with machine learning (ML) to both understand the availability of information relevant to LIHs on social media, and also to design automated machine learning tools for identifying such information, when available, in real time. In this way, we attempt to contribute societal crisis and ML fields in the following ways:

1. By examining what information was available during the pandemic, our findings can be used to further study how the information evolved during the period of the event, and also the implications that the patterns of public opinion and information may have for those attempting to manage a similar large-scale societal crisis.
2. The results of our combined content analysis and automated ML predictions on COVID-19 tweets provide insights that can be utilized to generate future recommendations for risk management and relief efforts for vulnerable populations. Specifically, our study can help with awareness, actionable items and policy recommendations for local utilities city planners, public housing authorities, landlords and relevant stakeholders.
3. The labeled dataset produced by our content analysis, together with ML models trained on carefully annotated data relevant to LIHs, can benefit the ML learning community focused

on social good applications, as well as communities focused on fairness and transparency in artificial intelligence.

5.2 Theoretical frameworks for content analysis

Content analysis is widely used in social science fields to analyze the systematic and replicable examination of symbols of communication. As opposed to the typical manual annotation, the traditional framework of content analysis consists of conceptualization, and purpose, design and analysis (Riffe et al., 2014; Zamith and Lewis, 2015). Current applications of content analysis have three distinct approaches: conventional, directed, or summative (Hsieh and Shannon, 2005). The major differences among the approaches are coding schemes, origins of codes, and threats to trustworthiness. In conventional content analysis or other text mining approaches, coding categories are generally derived directly from the text data. Based on the directed and summative approaches, however, this paper's analysis starts with a theory or relevant research findings as guidance for initial codes. As a conceptual approach, we utilize crisis communication frameworks and social order theory to better understand the social impacts of COVID-19 on LIHs through public opinions.

5.2.1 Crisis communication framework

To categorize public opinions expressed as tweets, this paper first adopts the framework of crisis communication that is widely utilized in the literature of crisis responses. Social media have become a community-based platform for crisis communication (Spence and Lachlan, 2016). Houston et al. (2015) developed a framework, outlining 15 different instances where social media can be utilized during and between the three final phases of a disaster: pre-event, event, and post-event. In this paper, we focus on the event and post-event phases. During the event, we aim to identify tweets that may be used to gain information about the pandemic, to deliver and receive news coverage, or to provide pandemic relief information, resources for mental health and support, well-wishes and concerns, and comments and discussions on personal or others' experiences. During the post-event, we aim to identify tweets that discuss the causes and implications of the pandemic, and emphasize

connections among community members.

5.2.2 Social order theory

Second, we use the social order theory to analyze the purpose of the tweets. More importantly, sociologists argue that crisis prevention, preparedness, response, and recovery can represent the multi-dimensions of social order (Kreps, 1984; Ritchie and MacDonald, 2010; Sovacool et al., 2020b). Overall, societal crises have a tremendous influence on the social order in the affected communities. According to Kreps (1989), social order can be characterized by specific entities that work to maintain social structure when it is in jeopardy (i.e., during a pandemic). These entities may be new or pre-existing, but they are shaped by human action. For example, Kreps (1989) states that, during a disaster, the police, hospital workers, or search and rescue personnel are all establishing and maintaining social order. Drawing on the elements from Simmel, Weber and Durkheim's interpretation of social order (Little et al., 2016), as well as previous disaster research, Kreps (1984) argues that forms of association enacted during disasters reflect sequences of four basic elements of social organization: domains (D), tasks (T), human and material resources (R), and activities (A). Based on this framework, we analyzed the tweets focusing on social-environmental impacts (human and material resources), status reporting, help requests, complaints, expressing thanks and well wishes, and general comments (activities).

5.2.3 Research questions

The primary goal of the content analysis is to investigate the nature of tweets relevant to LIHs generated within the time span of COVID-19 pandemic, in 2020. Specifically, we ask the following two research questions:

- To what extent did the low-income communities were affected by COVID-19 beyond the health impacts? What were the most tweeted categories related to low-income communities during the pandemic?
- To what extent are tweets during the pandemic about more than the reporting of COVID-19

status and health issues? What broader social, behavioral, economic and policy implications might be observed?

To answer these questions, we developed a procedure for creating content categories from Twitter, described below.

5.3 Procedure of content analysis

The purposes of using content analysis here are first to quantify, categorize and analyze the presence, meanings and relationships of certain keywords, themes, or concepts based on literature and theories; and to make inferences about the messages within the tweets through both qualitative and quantitative interpretations ([Carley, 1993](#)). Specifically, we developed the following key processes:

- the development of coding categories,
- the development of the coding protocol and sheet,
- the specification of the sample,
- the coder training for labeling tweets,
- the steps of discussing and resolving disagreements,
- the establishment of intercoder reliability.

5.3.1 Coding categories and coding protocol

To develop the coding categories, we used a multi-step approach. First, through a critical review of the literature and theories, we derived a list of content categories by considering the information involved in the event and post-event phases, including response and recovery. More specifically, the categories of the tweets were initially determined based on empirical COVID-19 studies and reports ([Chen et al., 2020a](#); [Gout and Kelly, 2020](#); [Graff and Carley, 2020](#); [Karpman et al., 2020](#); [Sovacool et al., 2020a](#)), disaster response literature, risk communication theories and an online source

CrisisLex.org concerning text analysis of natural disasters (Olteanu et al., 2014). CrisisLex.org is a repository of crisis related social media data and tools that many researchers studying disaster responses have widely adopted (Imran et al., 2015a; Nafi et al., 2020; Olteanu et al., 2015a). Second, through manual coding of Twitter content, we developed and revised other categories and sub-categories, and further examined the themes of each category to uncover public opinions and social issues during the pandemic. Using this approach, we derived a list of 15 categories based on the purpose of the tweets. Concurrently, we also developed the coding protocol and sheet, which contained the definitions of the categories, together with tweet examples for each category, as shown below:

1. **News and updates related to COVID-19 (NAU).** This category includes tweets that contain announcements, updates, information, policies and support programs. They should generally be a statement as opposed to an opinion. **Example:** *“House coronavirus bill aims to prevent utility shutoffs provides \$1.5 billion for states and Native tribes to help low-income households pay for drinking water and \$1.5 billion to help low-income households pay for energy. #NoShutoffs #EnergyJustice <https://t.co/FhWaBYxs86>”*
2. **Stay at home, quarantine (SHQ).** Tweets in this category contain reports about people that relocated/quarantined due to COVID-19, even for a short time, including experiences of others or self. **Example:** *“A billionaire somewhere bought an entire island a while ago and has relocated his family there during this period. There is food, internet, all the basic amenities you can think of. Imagine an island free of Covid-19. No one can go in or out. I watch too many movies.”*
3. **Food and medical supply shortage (FMS).** These are tweets that comment on supply shortage of PPE (personal protective equipments), sanitizing items (e.g., hand sanitizer), food, bottled water, and things alike. This category includes tweets from individuals, and organizations. Note: The category is about supply shortage, NOT about individuals that cannot afford them. **Example:** *“If we are having food, toilet paper, and medical supply shortages after a month and a half of a pandemic, I hope we can learn from our mistakes and stockpile supplies for the next one #coronavirus #COVID19”*

4. **Infected, hospitalized, and deaths (IHD).** Tweets that report or comment on casualties and/or infected people due to COVID-19 (including exposed, hospitalized, etc.), with focus on under-served communities.
5. **Complaints about service providers and landlords (CSP).** These are tweets that report issues with or complains about water/electricity/gas/internet services being disconnected; impaired living conditions (e.g., no A/C); and/or a lack of support for lower-income households (e.g., utility providers should not disconnect services and possibly waive LIH's bills.) The complaint has to target the utility provider or the landlord directly. **Example:** *Stopping utility shutoffs and offering financial assistance to low-income ratepayers is imperative in stopping the spread of the coronavirus and ensuring crucial access to water, electricity, internet, and more. #NoShutoffs*
6. **Policy responses (personal experiences and opinions toward specific policy) (PLR).** These are tweets that show supportive or opposite attitudes towards policies during COVID-19. This category includes comments on policies(e.g., statements about choosing not to wear a mask, violating quarantine, shutdown, staying at home orders) or comments on the implementation of a specific policy (e.g., stimulus package and check, assisting programs). This category also includes opinions, suggestions, and complaints toward government agencies and organizations under public regulations. **Example:** *#StimulusChecks #stimulus #Stimuluscheck #coronavirus #Food House Rent \$1800 Utilities \$475 Groceries "X" amount and the stimulus check \$1200? Intelligent step by Government. Can someone tell them that people need food to live as well? Note:Medical bill not included*
7. **Education and everyday life inconveniences or disruptions (ELI).** These are tweets that report, discuss, or complain about the impacts of COVID-19 on schools and the education system, and the impacts on other aspects of social life, such as business and recreation, that cause inconveniences (outside of the household). **Example:** *Inside life! Covid-19 has forced the education system to go digital. Many private schools are adjusting. But how effective can this be for the public school student without electricity, money and food. Or is the government*

expecting us to spend the little we have on generator?

8. **Income & economy impacts due to job loss or economics (IEI).** Tweets that report, discuss, or complain about the impacts of COVID-19 on personal incomes, or industry and economy, in general. Tweets about job loss and bankruptcy that affect low-income people belong to this category, including reports of difficulties in paying for foods, medicines, bills and rent. **Example:** *Things you might have missed. Tens of thousands of protestors in all 50 States. 40 Million unemployed unable to pay rent or mortgage WITH NO NATIONAL PLAN. 103K Dead from Covid-19 NO NATIONAL PLAN. WH lies about tear gassing its' own citizens.*
9. **COVID-19 Psychological impacts (CPI).** Reports or discussions about the psychological impacts of COVID-19, including direct and indirect impacts, e.g., worrying about getting infected, stress to get the ends meet, stress about working at home. **Example:** *Well, I need help bc I lost my job due to covid-19. I have no current income. I'm hurting really bad physically health wise in so many ways. I have taken blow after blow. I'm to weak to currently help myself. I wouldn't mind having something good happen to cheer me up.*
10. **Social inequality and justice issues (SIJ).** Tweets that report or discuss the social inequality and/or justice issues raised by, or indirectly caused by COVID-19, including exacerbated poverty, criminal, racial, and other issues. **Example:** *Coronavirus has taken the underlying issues with wealth inequality and access to internet and brought it to the forefront by forcing students from low income and rural areas to deal with e-learning without access to internet. Internet needs to be a public utility.*
11. **Caution and advice to general public (CAG).** Tweets that provide caution and/or advice for the general public. This category includes reports of warnings issued or lifted, guidance and tips. **Example:** *If you are receiving an income support payment that is eligible to receive the \$550 fortnightly Coronavirus supplement you will not be eligible for the second \$750 payment.*
12. **Donations & volunteering (DNV).** Raising awareness of urgent needs for resources (from the local community) or providing/donating supplies such as food, water, clothing, money,

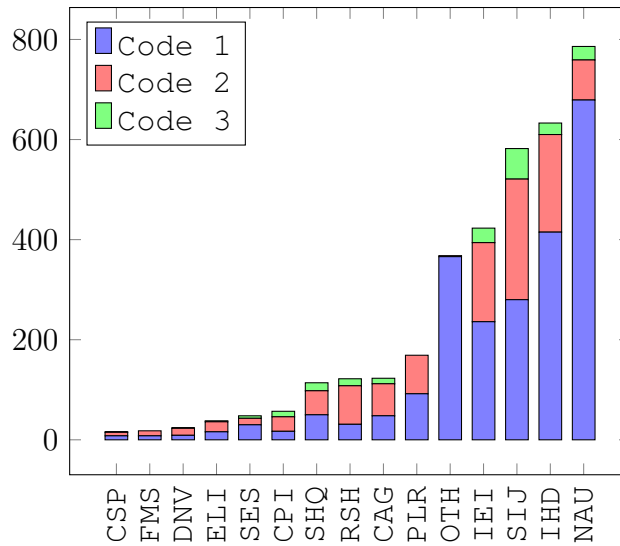


Figure 5.1: *Distribution over the categories provided by the coders in the manually annotated set.*

medical supplies, PPEs, or blood; also, raising awareness of volunteering services. Note: This does not include personal requests for help or donations.

13. **Sympathy & emotional support (SES).** Tweets that offer prayers, thoughts, and emotional support. **Example:** *Today we pray for the self-employed. Many have lost their income due to Covid 19 & are struggling to make ends meet. We pray for your provision & that they do not fall through the cracks of government support.#StayandPray*
14. **Requests for specific help, not comments (RSH)** Tweets that request help (for money, help on bill payment, essentials). A tweet in this category has to be calling for specific help, not comments. **Example:** *Well, I need help bc I lost my job due to covid-19. I have no current income. I'm hurting really bad physically health wise in so many ways. I have taken blow after blow. I'm to weak to currently help myself. I wouldn't mind having something good happen to cheer me up.*
15. **Others (OTH).** All the other COVID-19 LIH relevant tweets that do not belong to the categories listed above.

5.3.2 Sample data

We used Twitter Streaming API to crawl COVID-19 tweets that contain general keywords pertaining to COVID-19 (e.g. “coronavirus”, “covid-19”, “corona virus”, “#covid19”, etc.). The dataset crawled contained approximately 170 million tweets posted between March 23rd, 2020 and September 25th, 2020, and was stored in a MongoDB database. Using a carefully designed list of keywords relevant to LIHs (e.g., “low-income”, “poor”, “underserved communities”, “CARES Act”, “race”, etc.), we subsequently sampled tweets from the MongoDB database and constructed a dataset of tweets potentially useful for our LIH study. Random subsamples of the LIH dataset were used for coding. We ensured that the random samples contained only original English tweets (i.e., no retweets, comments, etc.) and no duplicate tweets.

5.3.3 Coding process and interoperability

We conducted five rounds of training for the coders, to ensure that the coders reached a satisfactory level of intercoder reliability, given the complexity of the categories considered. Each training lasted from one hour to 1.5 hours. Each of these rounds of training were aimed to ensure that (1) coders clearly understand the meanings of the categories; (2) coders clarify their thinking and coding process through discussions; (3) and that coders solve disagreement and reach consensus.

We used Cohen’s kappa to calculate intercoder reliability. One major shortcoming of using Cohen’s kappa to calculate intercoder reliability is that it assumes that one tweet only belongs to one category. However, in reality, one tweet may belong to more than one category. For example, a tweet can contain information about the user seeking help, while complaining about the utility companies. We allowed coders to categorize one tweet to maximum three categories, with the first choice as the primary choice. Cohen’s kappa was calculated based on the first choice. We also calculated fuzzy Kappa based on the procedures proposed in [Kirilenko and Stepchenkova \(2016\)](#). The fuzzy Kappa takes into account the exact situation where one subject may belong to multiple categories. However, fuzzy Kappa based on multiple categories was always greater than Cohen’s Kappa based on coders’ first choice through our calculation. As a result, we calculated both agreements, but chose to use Cohen’s Kappa to indicate intercoder reliability, as this is seen as

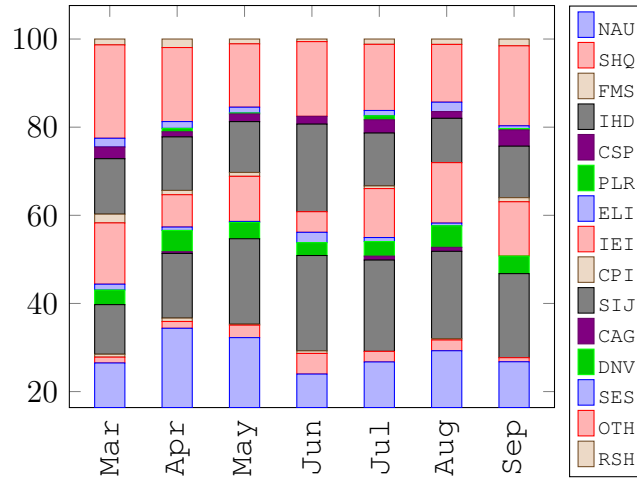


Figure 5.2: *Distribution of tweets over the months.*

a more conservative measure.

We had 6 coders in the first two training sessions, and the intercoder reliabilities based on Cohen's kappa were very low (the average of Cohen's Kappa among all possible pairs was 0.455). After the third training meeting, we trimmed the team of coders from 6 to 3, based on their performance. The rest of the three coders had two more trainings. Eventually, the three coders achieved great reliabilities based on the final coding sample that contained 112 tweets (the average of Cohen's Kappa among all possible pairs was 0.788, and Cohen's Kappa was .767, .863, .735 separately). We concluded the training upon great intercoder reliability.

5.3.4 Annotated dataset and insights provided

After the coder training was concluded, each of the three final coders independently annotated between 700 and 800 tweets, for a total of 2286 usable tweets (some tweets were not usable, e.g. if the coder used a wrong code). The distribution over the categories provided by the coders for each tweet is shown in Fig. 5.1, where the colors within a category correspond to the first code (blue), second code (pink) and third code (green), respectively, and the categories are ordered by the total number of tweets that contain that category. As expected, the NAU (news and updates) category is the largest in the annotated dataset, with almost 786 tweets, followed by the IHD (infected, hospitalized and deaths) category, focused on health impacts, with 633 tweets. The categories SIJ

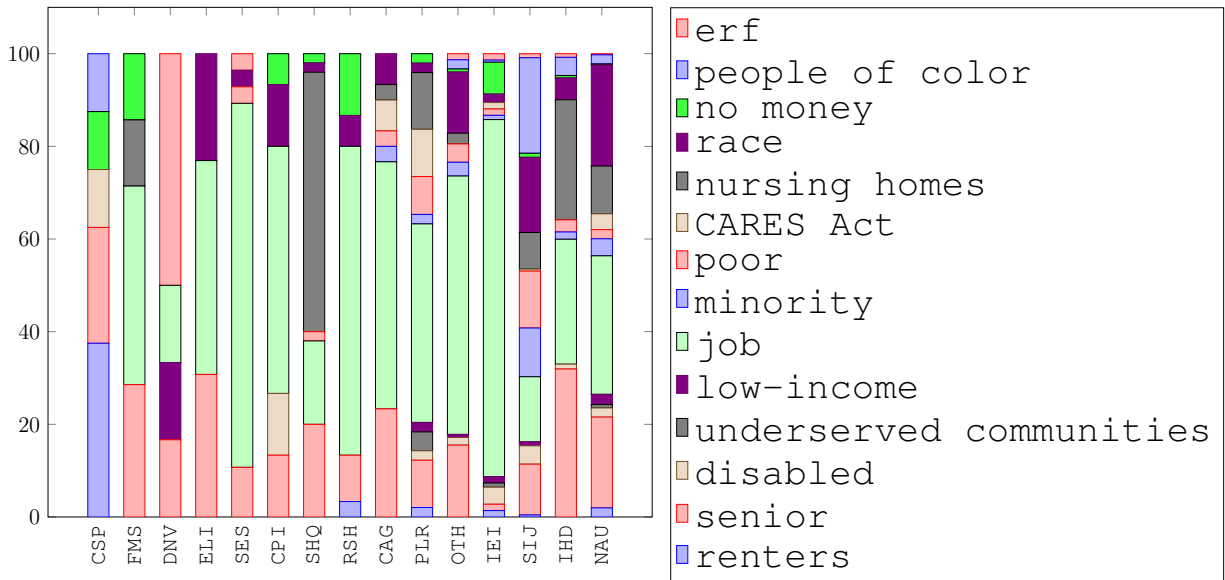


Figure 5.3: *Distribution over the keywords in the manually annotated set.*

(social inequality and justice) and IEI (income and economy impacts) have 582 and 423 tweets, respectively. The category PLR (policy responses) has 169 tweets, while the categories CAG (caution and advice), RSH (requests for specific help), and SHQ (stay home/quarantine) each have around 120 tweets. All the other categories have approximately 50 tweets or less, with the CSP (complains about service providers and landlords), FMS (food and medical supply shortage) and DNV (donations and volunteering) categories being the smallest with approximately 20 tweets each.

To gain insights into the evolution of the categories over time, Fig. 5.2 shows the category distribution in the tweets in our data sample over the months in which those tweets were crawled. As can be seen, the categories are similarly discussed over the months considered, although there are more news and updates (NAU) in April and May, more discussions about social inequality and justice issues (SIJ) in June, and more discussions about income and economy impacts in March, and then again later in July, August and September. To understand if the tweets in various categories can be identified using simple keyword searches, Fig. 5.3 shows the keyword distribution for the tweets in each of the categories considered in our coding scheme. As can be seen a mixture of keywords is present in each category, with some keywords such as senior, job, nursing homes, poor, people or color, race, etc. being prevalent in many categories. This makes it hard to identify a category based on keywords, and consequently, a keyword-based search approach cannot be used to accurately

extract specific LIH information from Twitter.

Although our data sample is relatively small, this content analysis shows that indeed social media can capture the effects of COVID-19 on LIHs, and in particular, it captures surprisingly well information about social inequality (SIJ), income and economy impacts (IEI), policy responses (PLR) and requests for specific help (RSH) relevant to LIHs. At the same time, it captures information about the physiological impacts of COVID-19 (CPI), education and everyday life inconveniences (ELI), complains about service providers and landlords (CSP), food and medical supply shortage (FMS), although these categories are encountered less frequently than others. The results of the content analysis, can thus be used to answer the research questions that we asked in Section 5.2.3. Specifically, the analysis suggests that while the health impacts of COVID-19 are the most prevalent, there are also significant impacts in terms of social inequalities, as well as income and economy losses, among others.

5.4 Automated ML models for identifying LIH information on social media

From a practical point of view, our analysis shows that information relevant to LIHs is available on social media and it can potentially be extracted in real time, if effective and efficient automated ML models are developed. Therefore, in this section, we explore the possibility of training accurate machine learning models for automatically extracting LIH information from Twitter.

5.4.1 Benchmark dataset for ML

Machine learning, in general, and deep-learning, in particular, require a relatively large amount of data to train a reliable classifier/model. The manual annotation of our data sample presented a significant imbalance in the category distribution as shown in Fig. 5.1, with some categories containing more than 600 instances, while others contain as few as 20 instances. Given the challenge of training classifiers to automatically recognize categories with a very small number of training instances, in the ML analysis presented in this section, we picked the five largest categories relevant

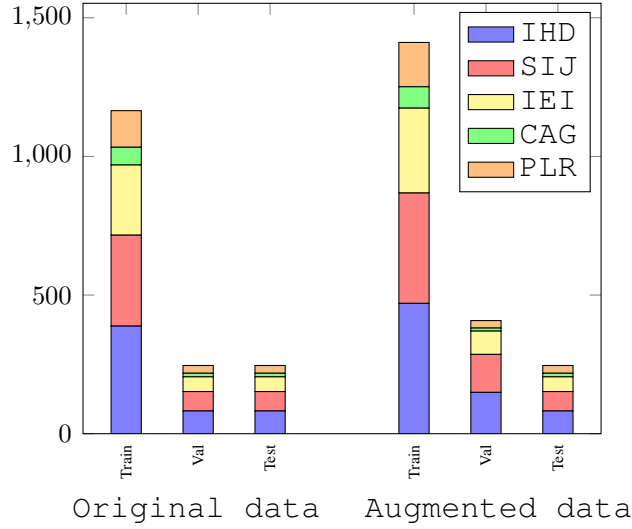


Figure 5.4: *Distribution of categories in the benchmark subsets, for the Original data (left) and the Augmented data (right).*

to LIHs, but ignored the NAU (news and updates) and OTH (Others) categories, which are broader and less interesting from the LIH point of view. Specifically, our ML classifiers are trained to identify the following categories: IHD (552 instances), SIJ (468 instances), IEI (359 instances), PLR (188 instances) and CAG (90 instances), for a total of 1657 instances. We should note that each tweet in the ML benchmark dataset belongs to exactly one category, although all three codes were considered when creating the dataset. However, priority was given to the first code, followed by the second code (if the first code was not among the five categories included in the ML analysis), and finally the third code (if the first two codes were not among the five categories of interest). The final dataset was split randomly into three subsets to be used for training (Train), validating (Val) and testing (Test), respectively. The category/class distribution over the Train, Val and Test subsets is shown in Fig. 5.4 (Original data).

5.4.2 Model design

To get an understanding of the ability of ML to accurately identify tweets into the five LIH categories considered, we used a typical deep learning classification architecture containing two main components: an embedding module and a classification head, as described below:

Embedding module

The embedding module is responsible for generating a condensed numeric representation of textual data. Formally, the embedding module is a mapping function $e_\theta : \mathbf{X} \rightarrow \hat{\mathbf{X}}$, where $X = \{x_1, \dots, x_n\}$ is the input text sequence and $\hat{X} = \{\hat{x}_1, \dots, \hat{x}_n\}$ is the contextual numeric representation of the tokens in the input text sequence.

In the realm of deep-learning for text classification, transformers (Vaswani et al., 2017) represent the most popular architectures used as embedding modules. Since the inception of the transformers, many pre-trained transformer-based language models have become available, and have helped achieve state-of-the-art results in various natural language processing (NLP) tasks (Devlin et al., 2018). Considering the success of transformer-based embedding models in NLP, in our study, we used the following pre-trained embedding models: BERT (Devlin et al., 2018), BERTweet (Nguyen et al., 2020), CovidBERT (Müller et al., 2020), BERT-corona (Karisani and Karisani, 2020), and BERTweet-covid. BERT is one of the first and most popular transformer-based pre-trained models, and was used as a baseline in our experiments. BERT is pre-trained using a large corpus of unlabeled texts including Wikipedia and Book Corpus (Devlin et al., 2018). The other four BERT variants were selected as they are pre-trained on tweets (BERTweet) and even COVID-19 tweets (CovidBERT, BERT-corona, and BERTweet-covid).

Classification head

The classification head uses the contextual representation of the input text sequence to predict a particular category for that input. Formally, the classification head is a mapping function $c_w : \hat{\mathbf{X}} \rightarrow \mathbf{y}$, where $y \in Y$, and Y is the set of all categories. The classification head in our model is a softmax dense layer with learnable parameters w .

5.4.3 Metrics

We used standard metrics like precision, recall and F1-measure to evaluate the performance of the models. However, considering the class-imbalance in the dataset, we also used weighted F1 along with macro-averaged to avoid unbalanced impact from minority under-performing or

Original ML Benchmark (Baseline Results)					
Model	Val F1	Pr	Re	F1	W-F1
BERT	76.73	68.17	66.29	67.21	72.67
BERTweet	76.54	74.57	70.78	72.63	76.41
CovidBERT	78.32	73.00	70.73	71.85	74.43
BERT-corona	78.03	73.24	68.14	70.60	75.70
BERTweet-covid	79.17	71.67	72.63	72.15	76.00
Augmented ML Benchmark (Human-in-the-loop Setting)					
Model	Val F1	Pr	Re	F1	W-F1
BERT	79.60	70.63	67.82	69.20	74.48
BERTweet	83.35	72.66	70.47	71.55	75.33
CovidBERT	82.10	73.63	69.46	71.48	74.44
BERT-corona	81.69	73.73	69.13	71.36	75.44
BERTweet-covid	83.70	77.23	73.00	75.06	78.51

Table 5.1: Average weighted $F1$ score ($W-F1$), precision, recall and macro average $F1$ scores for different experiments. The results shown are the average of five runs.

over-performing classes in the final score¹.

5.4.4 Experimental setup

We ran two sets of experiments, as described below:

- The goal of the first experiment was to get baseline results for the ability of the ML models to accurately identify the categories considered, with the limited amount of data annotated through content analysis. We used the benchmark dataset described in Section 5.4.1 to train and evaluate the BERT variants described in Section 5.4.2.
- The goal of the second experiment was to explore the potential improvement of the baseline models in a human-in-the-loop setting. To do that, we sampled 20,000 new unlabeled tweets from the MongoDB database using LIH-related keywords, just as the original data was sampled for coders. We then used the three BERT variants pre-trained on COVID-19 tweets to classify the 20,000 tweet in one of the five categories included in our analysis. Subsequently, we identified tweets for which the labels predicted by the three BERT models agreed, and

¹https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

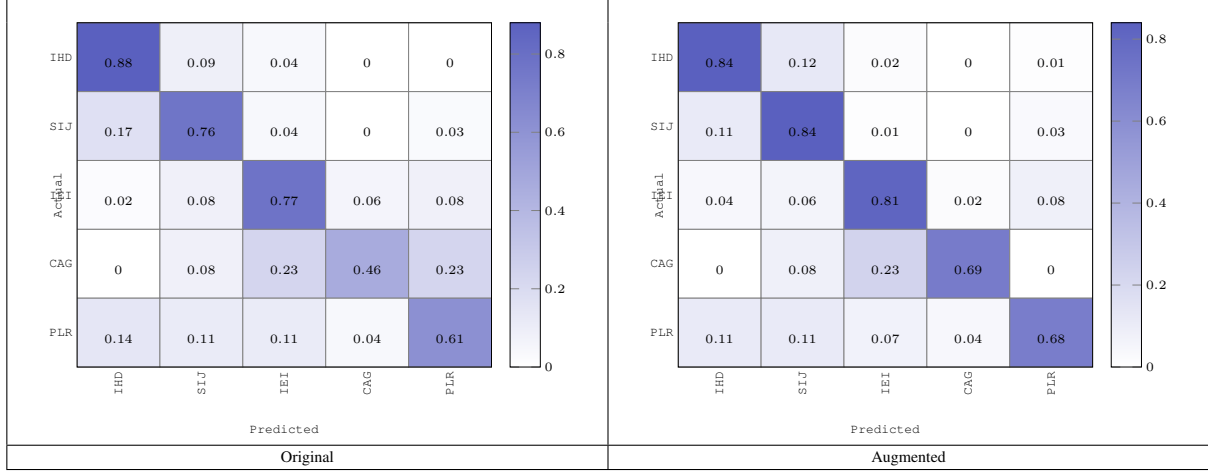


Table 5.2: *Confusion Matrices for the model’s prediction for original and augmented benchmark datasets.*

randomly sampled 600 of those tweets to be sent to the coders. We should note that the 20,000 tweets (and implicitly the 600 tweets selected) could potentially include tweets in all 15 categories used in the content analysis. However, our models were trained to identify only five categories (IHD, SIJ, IEI, PLR, CAG), and would thus assign one of these categories to each tweet in the set of 20,000 tweets. To avoid bias towards these categories and allow the coders to accurately label the new tweets, the original coding protocol was used, without telling the coders that the new set was biased towards a subset of the original categories. Out of the 600 tweets, 408 had at least one code in our set of categories. More specifically, the class distribution over the 408 tweets was as follows: IHD (149), SIJ (137), IEI (84), PLR (28), CAG (13). To ensure more variation in training, we added the original validation set to the training set, and used the newly labeled tweets as a new validation set. The test set was kept the same. We re-trained the models with the augmented training data. Figure 5.4 (Augmented data) shows the category/class distribution over the Train, Val and Test subsets is the augmented dataset.

The experiments were ran five times for each configuration and the results reported represent averages over the five runs.

5.5 ML results and discussion

Table 5.1 shows the results of ML model for the original and augmented dataset for different transformer variants. As can be seen in the table, the BERT baseline model is consistently outperformed by the domain-specific tweet-trained variants. Interestingly though, with the original benchmark, BERTweet performs the best overall (in terms of F1 and weighted F1), with BERTweet-covid in close second position. It should also be noted that BERTweet-covid increases recall, but decreases precision, as compared to BERTweet. However, more aligned with our hypothesis, BERTweet-covid significantly outperforms the other variants with the augmented benchmark dataset. This is most likely because the additional train and validation data in the augmented benchmark provide a better representation for the minority classes, and can thus benefit better from the COVID-specific embedding module. With the augmented benchmark, we can see improvement in performance in terms of weighted F1 when comparing the formal-text trained BERT with the general-tweet trained BERTweet and further with the COVID-specific variants BERT-corona and BERTweet-covid. Thus, the choice of the embedding module, particularly, the domain of the pre-training text, seems to be a deciding factor in model’s performance.

Moreover, in the human-in-the-loop setting, the augmented data together with BERTweet-covid provide the best overall result with a 3.3% W-F1 gain as compared to the standard BERT. It should be noted that with the augmentation, some variants show slightly lower performance, although the losses of those variants are not as significant as the gains of BERTweet-covid. Overall, the results show that a good choice of the transformer variant and human-in-the-loop data augmentation can help boost the performance.

Table 5.2 shows the confusion matrices for the best performing models for the original and augmented datasets, respectively. We can see that the data augmentation shows improvement of performance, especially for the minority classes (at a very small performance cost for the largest class). Most notably, the augmentation helps address the model’s confusion with the two smallest classes. For example, with the original data, 23% of the CAG tweets were classified as PLR tweets. Using the data augmentation, none of the CAG tweets were predicted as PLR. The confusion between these classes is probably caused by the fact that some CAG tweets provide advice on

No	Tweet	Label	Original Pred	Augmented Pred
1	If you are on a low income and have children, don't forget you may be entitled to FSM and be able to access a Food parcel or Vouchers.	CAG	PLR	CAG
2	Are you an employer who is having to furlough staff? Find out more about the Coronavirus Job Retention Scheme and which staff can be furloughed as part of the scheme.	CAG	PLR	CAG
3	In the Race for #coronavirus #Immunity, #Sweden leads the pack (The science is clear, immunity is real and Sweden is on its way to achieving herd immunity within the month. TS, Bill Gates?	CAG	PLR	CAG
4	During this time of national crisis, <USER>suggests limiting person to person contact until Coronavirus has been beaten. Play safe. Play online.	CAG	CAG	IEI

Table 5.3: Examples of tweets from the minority class CAG that were incorrectly classified by the models.

response policies. Table 5.3 shows three such example of tweets whose primary classification is CAG. The tweets 1, 2 and 3 were misclassified by the BERTweet-covid model trained with the original data, but correctly classified by the BERTweet-covid model trained with the augmented data. The additional labeled examples used for training were sufficient for the model to distinguish between policy response (PLR) tweet and caution and advice (CAG) tweets that mention policies. As can be seen in table 5.2 however, the model still incorrectly classifies 23% of CAG tweets as IEI even after data augmentation. Tweet 4 in Table 5.3 is an example of such tweets.

The numeric results in Table 5.1 together with the confusion matrices in Table 5.2 show that the combination of content analysis and machine learning has the potential to empower humanitarian organizations and responders with tools for identifying LIH information in real time.

5.6 Conclusions

In this chapter, we presented a framework which uses content analysis and machine learning to study the impact of COVID-19 on LIHs using social media. With content analysis, we proposed a coding strategy to annotate tweets and used the strategy to build a dataset relevant to LIH. Using ML, we trained ML models on this carefully annotated dataset to study the possibility of building

an automated tool for finding such content in real time.

Chapter 6

Concluding remarks and future prospects

Social media has been an integral part of modern life. The amount and variety of data shared in social media provide interesting clues towards understanding and solving important problems of the modern world. Complementary to that, with its recent advancements and growth, deep-learning presents itself as a powerful tool in making use of the unstructured data from social media. In this dissertation, I looked into the domain of crisis related social media posts and investigated various deep learning approaches in harnessing the potential of social media for the application of crisis response and management.

Firstly, we built a Twitter data collection system that makes use of the Twitter API to collect data in real-time. The system can be initialized with a set of seed keywords and will collect tweets for a desired topic or event with high precision. The system periodically updates the set of keywords to expand on the topics covered and increase recall. We used this system to collect data for various topics including COVID-19, hurricanes and other events.

After collecting tweets for various events, we created two manually annotated datasets for identifying fine-grained location mentions in crisis related tweets; one dataset contains tweets posted during several natural and man-made disasters and the other one tweets posted about COVID-19 pandemic. Identifying the location of concern mentioned in the text is integral in gaining situational awareness information about an emergent crisis. Location is considered one of the most important pieces of information for effective emergency response. We used crowd-sourcing to

annotate location phrases in the text with one of six categories: Country, State, Region/Continent, City/Town/County, Building/Landmark and Other. We also investigated various state-of-the-art architectures for the problem of location mention identification and established baseline results for the problem. We proposed a multi-task approach of using simplified auxiliary task of binary (location or not location) classification attached to some lower layer of the transformer model to improve performance on the main task of fine-grained location identification. With the two different datasets, we also explored cross-domain transferability between the models on zero-shot, embedding-level and model-level transfer. Finally, we performed error-analysis to understand the shortcomings of the model.

Building up on this work, we extended the investigation of multi-task learning to include datasets for other similar tasks for identifying fine-grained location mentions in tweets. We explored the effect of relatedness of tasks in the multi-task learning setting by co-training the model on several tasks, simultaneously. We used tasks like keyphrase identification, eyewitness-account classification and humanitarian categories classification as additional tasks for the main task of fine-grained location mention identification. Our results showed that using dissimilar tasks (i.e., location identification tasks together with other types of), yields better performance as compared to using similar tasks. This result is significant, especially, with respect to the time-critical nature of crisis events. For an emergent crisis, when only a small amount of data is available, using data for other tasks in a multi-task setting could be a time-saving and better performing approach than annotating more data for the same task.

Next, we used contrastive learning for multimodal classification of crisis related tweets. In particular, we used the contrastive language-image pretraining (CLIP) algorithm to pre-train transformer pairs for the image and text modalities. We showed that CLIP pre-trained models perform better for the task of *Informativeness* and *Humanitarian* categories classification from the *CrisisMMD* dataset. We also showed that domain-specific pre-training helps improve the performance, especially for the more complex task of *Humanitarian* categories classification which has a smaller amount of labelled data. We also performed few-shot experiments and showed that the CLIP pre-trained models achieve better score than baseline models with only a small fraction of the total data.

In the final chapter, we used social media to analyze the public response to COVID-19 pandemic.

In particular, we investigated ways to identify information relevant to the impact of the pandemic on the low-income population. We used content analysis based on crisis communication framework and social order theory to build categories for different topics relevant to low-income households. We manually annotated tweets to create a dataset and trained deep learning models to automatically identify such tweets. We experimented with different variants of pre-trained transformer based models. We also employed a human-in-the-loop approach as data-augmentation. Our results showed that the human-in-the-loop method helps improve the performance, especially, for the less represented classes.

As part of future work, the location identification task can be used for linking the locations identified to precise geo-locations with geo-coding. The information can then be used to build dashboards which can be used by the emergency responders to identify areas of interest. For multi-task learning, we plan to explore the co-training at a larger scale with a larger number of tasks. We currently only looked into the performance of the main task in the multi-task settings. As an extension, we plan to study the effect of multi-task learning on all the tasks involved. Moreover, it would also be interesting to look into the prospect of transfer learning between the two location identification tasks. For the task of contrastive learning, we plan to dive deep into the multimodal representations learned for similar and dissimilar text-image pairs. The insight into the representations of pairs with differing parity as compared to the similar pairs can be useful in improving the representations. For the task of identifying information relevant to low-income households, we noted that the imbalance in the data restricts from building a full-fledged ML model for all the classes derived using content analysis. As part of our future work, we plan to explore ways to build models for all the derived classes using ML strategies such as self-supervised learning and knowledge distillation.

Bibliography

- M. Abavisani, L. Wu, S. Hu, J. Tetreault, and A. Jaimes. Multimodal categorization of crisis events in social media. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14679–14689, 2020.
- O. Adjali, R. Besançon, O. Ferret, H. L. Borgne, and B. Grau. Multimodal entity linking for tweets. In European Conference on Information Retrieval, pages 463–478. Springer, 2020.
- F. Alam, S. Joty, and M. Imran. Domain adaptation with adversarial training and graph embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018a.
- F. Alam, F. Ofli, and M. Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018, pages 465–473. AAAI Press, 2018b.
- A. Anand, A. Awekar, et al. Fine-grained entity type classification by jointly learning representations and label embeddings. arXiv preprint arXiv:1702.06709, 2017.
- S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015.
- R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4):555–596, 2008.
- A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, and M. Auli. Cloze-driven pretraining of self-attention networks. arXiv preprint arXiv:1903.07785, 2019.

- D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
- J. G. A. Barbedo. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Computers and electronics in agriculture, 153: 46–53, 2018.
- V. Beard. Covid-19: Poverty, housing, homelessness—a broad view and a picture from west michigan. Housing, Homelessness—A Broad View and a Picture from West Michigan (May 28, 2020), 2020.
- E. A. Benfer, S. Mohapatra, L. F. Wiley, and R. Yearby. Health justice strategies to combat the pandemic: Eliminating discrimination, poverty, and health disparities during and after covid-19. Yale Journal of Health Policy, Law, and Ethics, 19(3):3, 2020.
- R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research, 55:409–442, 2016.
- G. Biamby, G. Luo, T. Darrell, and A. Rohrbach. Twitter-comms: Detecting climate, covid, and military multimodal misinformation. arXiv preprint arXiv:2112.08594, 2021.
- R. Bielawski, B. Devillers, T. Van De Cruys, and R. Vanrullen. When does CLIP generalize better than unimodal models? when judging human-centric concepts. In Proceedings of the 7th Workshop on Representation Learning for NLP, pages 29–38, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.4. URL <https://aclanthology.org/2022.repl4nlp-1.4>.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.
- H. Cai, Y. Yang, X. Li, and Z. Huang. What are popular: exploring twitter features for event detection, tracking and visualization. In Proceedings of the 23rd ACM international conference on Multimedia, pages 89–98, 2015.

- K. Carley. Coding choices for textual analysis: A comparison of content analysis and map analysis. Sociological methodology, pages 75–126, 1993.
- R. Caruana. Multitask learning. Machine Learning, 28(1):41–75, Jul 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734. URL <https://doi.org/10.1023/A:1007379606734>.
- Y. Charalabidis and E. Loukis. Participative public policy making through multiple social media platforms utilization. International Journal of Electronic Government Research (IJEGR), 8(3): 78–97, 2012.
- C.-f. Chen, G. Z. de Rubens, X. Xu, and J. Li. Coronavirus comes home? energy use, home energy management, and the social-psychological factors of covid-19. Energy research & social science, 68:101688, 2020a.
- Q. Chen, W. Wang, K. Huang, S. De, and F. Coenen. Multi-modal adversarial training for crisis-related data classification on social media. In 2020 IEEE International Conference on Smart Computing (SMARTCOMP), pages 232–237. IEEE, 2020b.
- S. Chen and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pages 49–56, 2015.
- B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan. Multi-task low-rank affinity pursuit for image segmentation. In 2011 International Conference on Computer Vision, pages 2439–2446. IEEE, 2011.
- J. R. Chowdhury, C. Caragea, and D. Caragea. On identifying hashtags in disaster twitter data. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 498–506, 2020.
- J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46, 1960.

- M. Dahab, K. Van Zandvoort, S. Flasche, A. Warsame, R. Ratnayake, C. Favas, P. B. Spiegel, R. J. Waldman, and F. Checchi. Covid-19 control in low-income settings and displaced populations: what can realistically be done? Conflict and health, 14(1):1–6, 2020.
- S. Das and A. K. Kolya. Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on covid-19 by deep convolutional neural network. Evolutionary Intelligence, pages 1–22, 2021.
- A. Deaton. Covid-19 and global income inequality. Technical report, National Bureau of Economic Research, 2021.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.
- J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.
- L. Duong, T. Cohn, S. Bird, and P. Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 845–850, 2015.
- J. Elster. The cement of society: A survey of social order. Cambridge university press, 1989.
- J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- V. Galasso. Covid: Not a great equalizer. CESifo Economic Studies, 66(4):376–393, 2020.
- A. K. Gautam, L. Misra, A. Kumar, K. Misra, S. Aggarwal, and R. R. Shah. Multimodal analysis of

- disaster tweets. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), pages 94–103. IEEE, 2019.
- K. Glandt, S. Khanal, Y. Li, D. Caragea, and C. Caragea. Stance detection in covid-19 tweets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1596–1611, 2021.
- X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh and M. Titterington, editors, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, volume 9 of Proceedings of Machine Learning Research, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.
- E. Gout and C. Kelly. Extreme heat during the covid-19 pandemic amplifies racial and economic inequities, 2020. Online; posted June 29, 2020.
- A. Goyal, V. Gupta, and M. Kumar. Recent named entity recognition and classification techniques: a systematic review. Computer Science Review, 29, 2018.
- M. Graff and S. Carley. Covid-19 assistance needs to target energy insecurity. Nature Energy, 5(5): 352–354, 2020.
- T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen. Cooperative multimodal approach to depression detection in twitter. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 110–117, 2019.
- B. Han, A. J. Yepes, A. MacKinlay, and Q. Chen. Identifying twitter location mentions. In Proceedings of the Australasian Language Technology Association Workshop 2014, 2014.
- X. Han and J. Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4238–4248, Hong Kong, China, Nov. 2019. Association

- for Computational Linguistics. doi: 10.18653/v1/D19-1433. URL <https://www.aclweb.org/anthology/D19-1433>.
- T. Harris and J. Lewis. Energy insecurity, energy justice & equity measurements, low-income weatherization & energy efficiency programming: National & maryland policies. 2020.
- A. Hekler, J. S. Utikal, A. H. Enk, W. Solass, M. Schmitt, J. Klode, D. Schadendorf, W. Sondermann, C. Franklin, F. Bestvater, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. European Journal of Cancer, 118:91–96, 2019.
- T. B. N. Hoang, V. Moriceau, and J. Mothe. Predicting locations in tweets. In Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CLICLing 2017), 2017.
- S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9, 1997.
- J. B. Houston, J. Hawthorne, M. F. Perreault, E. H. Park, M. Goldstein Hode, M. R. Halliwell, S. E. Turner McGowen, R. Davis, S. Vaid, J. A. McElderry, et al. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. Disasters, 39(1):1–22, 2015.
- H.-F. Hsieh and S. E. Shannon. Three approaches to qualitative content analysis. Qualitative health research, 15(9):1277–1288, 2005.
- K. Hunt, B. Wang, and J. Zhuang. Misinformation debunking and cross-platform information sharing through twitter during hurricanes harvey and irma: a case study on shelters and id checks. Natural Hazards, 103(1):861–883, 2020.
- K. Hunt, P. Agarwal, and J. Zhuang. Monitoring misinformation on twitter during crisis events: a machine learning approach. Risk analysis, 42(8):1728–1748, 2022.
- Y. Ikawa, M. Vukovic, J. Rogstadius, and A. Murakami. Location-based insights from the social web. In Proceedings of the 22nd international conference on World Wide Web, 2013.

- A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. IEEE Access, 8: 181074–181090, 2020.
- M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. ACM Comput. Surv., 47(4), June 2015a. ISSN 0360-0300. doi: 10.1145/2771588. URL <https://doi.org/10.1145/2771588>.
- M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR), 47, 2015b.
- M. Imran, P. Mitra, and C. Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.
- D. Inkpen, J. Liu, A. Farzindar, F. Kazemi, and D. Ghazi. Detecting and disambiguating locations mentioned in twitter messages. In A. Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, pages 321–332, Cham, 2015. Springer International Publishing.
- U. Irfan and B. Resnick. Megadisasters devastated america in 2017. and they’re only going to get worse., Mar 2018. URL <https://www.vox.com/energy-and-environment/2017/12/28/16795490/natural-disasters-2017-hurricanes-wildfires-heat-climate-change-cost-deaths>.
- Z. Ji, A. Sun, G. Cong, and J. Han. Joint recognition and linking of fine-grained locations from tweets. In Proceedings of the 25th International Conference on World Wide Web, 2016.
- Y. Jiang, C. Hu, T. Xiao, C. Zhang, and J. Zhu. Improved differentiable architecture search for language modeling and named entity recognition. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3585–3590, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

- L. Kaewkitipong, C. Chen, and P. Ractham. Lessons learned from the use of social media in combating a crisis: A case study of 2011 thailand flooding disaster. 2012.
- K. Kang, C. Yoon, and E. Y. Kim. Identifying depressive users in twitter using multimodal analysis. In 2016 International Conference on Big Data and Smart Computing (BigComp), pages 231–238. IEEE, 2016.
- N. Karisani and P. Karisani. Mining coronavirus (covid-19) posts in social media, Jul 2020. URL https://figshare.com/articles/dataset/Mining_Coronavirus_COVID-19_Posts_in_Social_Media/12597755/1.
- M. Karpman, S. Zuckerman, D. Gonzalez, and G. M. Kenney. The covid-19 pandemic is straining families’ abilities to afford basic needs. Washington, DC: Urban Institute, 500, 2020.
- S. Khanal and D. Caragea. Multi-task learning to enable location mention identification in the early hours of a crisis event. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4051–4056, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.340. URL <https://aclanthology.org/2021.findings-emnlp.340>.
- S. Khanal, R. Refati, K. Glandt, D. Caragea, S. Xu, and C.-f. Chen. Using content analysis and machine learning to identify covid-19 information relevant to low-income households on social media. In 2021 IEEE Intl Conf on Parallel Distributed Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking (ISPA/ BDCloud/ SocialCom/ SustainCom), pages 1522–1531, 2021. doi: 10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00205.
- S. Khanal, M. Traskowsky, and D. Caragea. Identification of fine-grained location mentions in crisis tweets. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 7164–7173, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.776>.

- L. King. Social media use during natural disasters: An analysis of social media usage during hurricanes harvey and irma. In Proceedings of the International Crisis and Risk Communication Conference, volume 1, pages 20–23, 03 2018.
- A. P. Kirilenko and S. Stepchenkova. Inter-coder agreement in one-to-many classification: fuzzy kappa. PloS one, 11(3):e0149787, 2016.
- G. A. Kreps. Sociological inquiry and disaster research. Annual review of sociology, 10(1): 309–330, 1984.
- G. A. Kreps. Social structure and disaster; symposium on social structure and disaster, college of William and Mary, Williamsburg, Virginia, 15-16 may 1986. University of Delaware Press, 1989.
- M. Kreuter, R. Garg, I. Javed, B. Golla, J. Wolff, and C. Charles. 3.5 million social needs requests during covid-19: What can we learn from 2–1-1? Health Affairs, 2020.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012.
- J. Kropczynski, R. Grace, J. Coche, S. Halse, E. Obeysekare, A. Montarnal, F. Benaben, and A. Tapia. Identifying actionable information on social media for emergency dispatch. Proceedings of the ISCRAM Asia Pacific, page 11, 2018.
- A. Kumar and G. Garg. Sentiment analysis of multimodal twitter data. Multimedia Tools and Applications, 78(17):24103–24119, 2019.
- A. Kumar and J. P. Singh. Location reference identification from tweets during emergencies: A deep learning approach. International journal of disaster risk reduction, 33, 2019.
- J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, 2001.

- A. Lal et al. Sane 2.0: System for fine grained named entity typing on textual data. Engineering Applications of Artificial Intelligence, 84, 2019.
- Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.
- C. Li and A. Sun. Fine-grained location extraction from tweets with temporal awareness. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, page 43–52, New York, NY, USA, 2014. Association for Computing Machinery.
- J. Li, A. Sun, J. Han, and C. Li. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 2020.
- J. Lingad, S. Karimi, and J. Yin. Location extraction from disaster-related microblogs. In Proceedings of the 22nd International Conference on World Wide Web, WWW '13 Companion, page 1017–1020, New York, NY, USA, 2013. Association for Computing Machinery.
- W. Little, R. McGivern, and N. Kerins. Introduction to Sociology-2nd Canadian Edition. BC Campus, 2016.
- F. Liu, M. Vasardani, and T. Baldwin. Automatic identification of locative expressions from social media text: A comparative analysis. In Proceedings of the 4th International Workshop on Location and the Web, pages 9–16, 2014.
- X. Liu, P. He, W. Chen, and J. Gao. Multi-task deep neural networks for natural language understanding. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4487–4496, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1441. URL <https://www.aclweb.org/anthology/P19-1441>.
- Y. Liu, F. Meng, J. Zhang, J. Xu, Y. Chen, and J. Zhou. GCDT: A global context enhanced deep transition architecture for sequence labeling. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, July 2019b.

- Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019c.
- I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.
- Y. Luo, F. Xiao, and H. Zhao. Hierarchical contextualized representation for named entity recognition. arXiv preprint arXiv:1911.02257, 2019.
- A. L’heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz. Machine learning with big data: Challenges and approaches. Ieee Access, 5:7776–7797, 2017.
- X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1064–1074, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.
- S. Madichetty, S. Muthukumarasamy, and P. Jayadev. Multi-modal classification of twitter data during disasters for humanitarian response. Journal of Ambient Intelligence and Humanized Computing, 12(11):10223–10237, 2021.
- S. Madichetty et al. Classifying informative and non-informative tweets from the twitter by adapting image features during disaster. Multimedia Tools and Applications, 79(39):28901–28923, 2020.
- A. Magge, D. Weissenbacher, A. Sarker, M. Scotch, and G. Gonzalez-Hernandez. Bi-directional recurrent neural network models for geographic location extraction in biomedical literature. In PSB, 2019.
- J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? inferring home locations of twitter users. In Sixth International AAAI Conference on Weblogs and Social Media, 2012.
- S. Malmasi and M. Dras. Location mention detection in tweets and microblogs. In Conference of the Pacific Association for Computational Linguistics. Springer, 2015.

- I. Mani, C. Doran, D. Harris, J. Hitzeman, R. Quimby, J. Richer, B. Wellner, S. Mardis, and S. Clancy. Spatiaiml: annotation scheme, resources, and evaluation. Language Resources and Evaluation, 44(3):263–280, 2010.
- S. E. Middleton, L. Middleton, and S. Modafferi. Real-time crisis mapping of natural disasters using social media. IEEE Intelligent Systems, 29(2):9–17, June 2014. URL <https://eprints.soton.ac.uk/370581/>.
- M. Müller, M. Salathé, and P. E. Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503, 2020.
- N. M. Nafi, A. Bose, S. Khanal, D. Caragea, and W. H. Hsu. Abstractive text summarization of disaster-related document. In ISCRAM 2020 Conference Proceedings–17th International Conference on Information Systems for Crisis Response and Management, 2020.
- G. Nalluru, R. Pandey, and H. Purohit. Relevancy classification of multimodal social media streams for emergency services. In 2019 IEEE International Conference on Smart Computing (SMARTCOMP), pages 121–125. IEEE, 2019.
- L. Nemes and A. Kiss. Social media sentiment analysis based on covid-19. Journal of Information and Telecommunication, 5(1):1–15, 2021.
- J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In ICML, pages 689–696, 2011. URL https://icml.cc/2011/papers/399_icmlpaper.pdf.
- D. Q. Nguyen, T. Vu, and A. T. Nguyen. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020.
- D. T. Nguyen, K. A. Al Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In ICWSM, 2017.
- F. Ofli, F. Alam, and M. Imran. Analysis of social media data using multimodal deep learning for disaster response. arXiv preprint arXiv:2004.11838, 2020.

- A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In Proceedings of the International AAAI Conference on Web and Social Media, volume 8, 2014.
- A. Olteanu, S. Vieweg, and C. Castillo. What to expect when the unexpected happens: Social media communications across crises. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15, page 994–1009, New York, NY, USA, 2015a. Association for Computing Machinery. ISBN 9781450329224. doi: 10.1145/2675133.2675242. URL <https://doi.org/10.1145/2675133.2675242>.
- A. Olteanu, S. Vieweg, and C. Castillo. What to expect when the unexpected happens: Social media communications across crises. In CSCW, 2015b.
- T. Oyedare and J.-M. J. Park. Estimating the required training dataset size for transmitter classification using deep learning. In 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), pages 1–10. IEEE, 2019.
- K. Parker, R. Minkin, and J. Bennett. Economic fallout from covid-19 continues to hit lower-income americans the hardest. Pew Research Center, page 21, 2020.
- J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014.
- U. Qazi, M. Imran, and F. Ofli. Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information. arXiv preprint arXiv:2005.11177, 2020.
- A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.
- A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.

- N. Rank, B. Pfahringer, J. Kempfert, C. Stamm, T. Kühne, F. Schoenrath, V. Falk, C. Eickhoff, and A. Meyer. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. NPJ digital medicine, 3(1):1–12, 2020.
- J. Rao, F. Ture, and J. Lin. Multi-task learning with neural networks for voice query understanding on an entertainment platform. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 636–645, 2018.
- M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.
- D. Riffe, S. Lacy, and F. Fico. Analyzing Media Messages: Using Quantitative Content Analysis in Research. Routledge Studies in Contemporary Philosophy. Taylor & Francis, 2014. ISBN 9781135912086. URL https://books.google.com/books?id=am_MAgAAQBAJ.
- L. A. Ritchie and W. MacDonald. Enhancing disaster and emergency preparedness, response, and recovery through evaluation. New Directions for Evaluation, 2010(126):3–7, 2010.
- R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- S. Ruder. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098, 2017.
- R. Safa, P. Bayat, and L. Moghtader. Automatic detection of depression symptoms in twitter using multimodal analysis. The Journal of Supercomputing, 78(4):4709–4744, 2022.
- C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL <https://arxiv.org/abs/2205.11487>.

- T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, pages 851–860, 2010.
- T. J. Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. Proceedings of the National Academy of Sciences, 117(48):30033–30038, 2020.
- H. D. Sherali, J. Desai, and T. S. Glickman. Allocating emergency response resources to minimize risk with equity considerations. American Journal of Mathematical and Management Sciences, 24(3-4):367–410, 2004. doi: 10.1080/01966324.2004.10737638. URL <https://doi.org/10.1080/01966324.2004.10737638>.
- D. Shome and T. Kar. Conoffense: Multi-modal multitask contrastive learning for offensive content identification. In 2021 IEEE International Conference on Big Data (Big Data), pages 4524–4529. IEEE, 2021.
- K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. URL <https://arxiv.org/abs/1409.1556>.
- R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, Advances in Neural Information Processing Systems, volume 26. Curran Associates, Inc., 2013. URL <https://proceedings.neurips.cc/paper/2013/file/2d6cc4b2d139a53512fb8cbb3086ae2e-Paper.pdf>.
- T. Sosea, I. Sirbu, C. Caragea, D. Caragea, and T. Rebedea. Using the image-text relationship to improve multimodal disaster tweet classification. In The 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021), 2021.
- B. K. Sovacool, D. F. Del Rio, and S. Griffiths. Contextualizing the covid-19 pandemic for a carbon-constrained world: Insights for sustainability transitions, energy justice, and research methodology. Energy Research & Social Science, 68:101701, 2020a.

- B. K. Sovacool, X. Xu, G. Zarazua De Rubens, and C.-F. Chen. Social media and disasters: human security, environmental racism, and crisis communication in hurricane irma response. Environmental Sociology, 6(3):291–306, 2020b.
- P. R. Spence and K. A. Lachlan. Reoccurring challenges and emerging threats. The handbook of international crisis communication research, 43:212, 2016.
- R. Stiegler, S. Tilley, and T. Parveen. Finding family and friends in the aftermath of a disaster using federated queries on social networks and websites. In 2011 13th IEEE International Symposium on Web Systems Evolution (WSE), pages 21–26. IEEE, 2011.
- D. Stoller, S. Ewert, and S. Dixon. Jointly detecting and separating singing voice: A multi-task approach. In International Conference on Latent Variable Analysis and Signal Separation, pages 329–339. Springer, 2018.
- K. Stowe, M. Paul, M. Palmer, L. Palen, and K. M. Anderson. Identifying and categorizing disaster-related tweets. In Proceedings of The fourth international workshop on natural language processing for social media, pages 1–6, 2016.
- C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision, pages 843–852, 2017.
- H. To, S. Agrawal, S. H. Kim, and C. Shahabi. On identifying disaster-related tweets: Matching-based or learning-based? In 2017 IEEE third international conference on multimedia big data (BigMM), pages 330–337. IEEE, 2017.
- S.-F. Tsao, H. Chen, T. Tisseverasinghe, Y. Yang, L. Li, and Z. A. Butt. What social media told about us in the time of covid-19: a scoping review. The Lancet Digital Health, 3:e175 – e194, 2021.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

- A. Vempala and D. Preotiuc-Pietro. Categorizing and inferring the relationship between the text and image of twitter posts. In Proceedings of the 57th annual meeting of the Association for Computational Linguistics, pages 2830–2840, 2019.
- S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In Proceedings of the SIGCHI conference on human factors in computing systems, pages 1079–1088, 2010.
- A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL <https://www.aclweb.org/anthology/W18-5446>.
- J. Wang, L. Sun, Y. Liu, M. Shao, and Z. Zheng. Multimodal sarcasm target identification in tweets. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8164–8175, 2022.
- T. Wang, Y. Tao, S. C. Chen, and M. L. Shyu. Multi-task multimodal learning for disaster situation assessment. In 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pages 209–212, 2020. doi: 10.1109/MIPR49039.2020.00050.
- Y. Wang, M. McKee, A. Torbica, and D. Stuckler. Systematic literature review on the spread of health-related misinformation on social media. Social Science & Medicine, 240:112552, 2019. ISSN 0277-9536. doi: <https://doi.org/10.1016/j.socscimed.2019.112552>. URL <https://www.sciencedirect.com/science/article/pii/S0277953619305465>.
- C. Wendling, J. Radisch, and S. Jacobzone. The use of social media in risk and crisis communication. (24), 2013. doi: <https://doi.org/https://doi.org/10.1787/5k3v01fskp9s-en>. URL <https://www.oecd-ilibrary.org/content/paper/5k3v01fskp9s-en>.
- P. Wicke and M. M. Bolognesi. Covid-19 discourse on twitter: How the topics, sentiments, subjectivity, and figurative frames changed over time. Frontiers in Communication, 6:45, 2021.

- T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface’s transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771, 2019.
- M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In Proc. INTERSPEECH 2010, Makuhari, Japan, pages 2362–2365, 2010.
- Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google’s neural machine translation system: Bridging the gap between human and machine translation. CoRR, abs/1609.08144, 2016.
- C. Xu, J. Li, X. Luo, J. Pei, C. Li, and D. Ji. Dlocrl: A deep learning pipeline for fine-grained location recognition and linking in tweets. In The World Wide Web Conference, 2019.
- L. Xukun and D. Caragea. Improving disaster-related tweet classification with a multimodal approach. In ISCRAM 2020 Conference Proceedings–17th International Conference on Information Systems for Crisis Response and Management, 2020.
- I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6442–6454, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.523. URL <https://aclanthology.org/2020.emnlp-main.523>.
- Y. Yang and T. M. Hospedales. Trace norm regularised deep multi-task learning. arXiv preprint arXiv:1606.04038, 2016.
- Y. Yilmaz and A. O. Hero. Multimodal event detection in twitter hashtag networks. Journal of Signal Processing Systems, 90(2):185–200, 2018.

- J. Yu and J. Jiang. Adapting bert for target-oriented multimodal sentiment classification. *IJCAI*, 2019.
- K. Zahra, M. Imran, and F. O. Ostermann. Automatic identification of eyewitness messages on twitter during disasters. *Information Processing & Management*, 57(1):102107, 2020. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2019.102107>. URL <http://www.sciencedirect.com/science/article/pii/S0306457319303590>.
- R. Zamith and S. C. Lewis. Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The ANNALS of the American Academy of Political and Social Science*, 659(1):307–318, 2015.
- Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 94–108, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.
- B. Zhu, X. Zheng, H. Liu, J. Li, and P. Wang. Analysis of spatiotemporal characteristics of big data on social media sentiment with covid-19 epidemic topics. *Chaos, Solitons & Fractals*, 140: 110123, 2020.
- Z. Zou, H. Gan, Q. Huang, T. Cai, and K. Cao. Disaster image classification by fusing multimodal social media data. *ISPRS International Journal of Geo-Information*, 10(10):636, 2021.