A Comparison Study for Disaster Tweet Classification Using Deep Learning Models

Soudabeh Taghian Dinani and Doina Caragea
Kansas State University, Manhattan, Kansas, U.S.A.

Keywords: Tweet Classification, Capsule Neural Networks, BERT, LSTM, Bi-LSTM.

Abstract: Effectively filtering and categorizing the large volume of user-generated content on social media during disaster events can help emergency management and disaster response prioritize their resources. Deep learning approaches, including recurrent neural networks and transformer-based models, have been previously used for this purpose. Capsule Neural Networks (CapsNets), initially proposed for image classification, have been proven to be useful for text analysis as well. However, to the best of our knowledge, CapsNets have not been used for classifying crisis-related messages, and have not been extensively compared with state-of-the-art transformer-based models, such as BERT. Therefore, in this study, we performed a thorough comparison between CapsNet models, state-of-the-art BERT models and two popular recurrent neural network models that have been successfully used for tweet classification, specifically, LSTM and Bi-LSTM models, on the task of classifying crisis tweets both in terms of their informativeness (binary classification), as well as their humanitarian content (multi-class classification). For this purpose, we used several benchmark datasets for crisis tweet classification, namely CrisisBench, CrisisNLP and CrisisLex. Experimental results show that the performance of the CapsNet models is on a par with that of LSTM and Bi-LSTM models for all metrics considered, while the performance obtained with BERT models have surpassed the performance of the other three models across different datasets and classes for both classification tasks, and thus BERT could be considered the best overall model for classifying crisis tweets.

1 INTRODUCTION

Social media plays a significant role in many people's lives and has changed the way people communicate virtually (Young et al., 2020). Being a faster way of news distribution and providing a two-way communication channel, social media has been increasingly gaining popularity and has grown into a main medium of human interaction and communication (Young et al., 2020). Therefore, it is not surprising that many people turn to social media during disaster events. (such as earthquakes, hurricanes, fires, etc.). They use social media platforms to connect with family and friends, to seek help and emotional support, to find information about food, shelter and transportation and to share and distribute information and news (Ahmed, 2011), especially as standard communication mediums (such as 911 lines) may be unavailable due to the large number of calls received during disasters (Villegas et al., 2018). For example, during the floods caused by Hurricane Harvey in Texas in August 2017, not being able to get through emergency lines, a lot of trapped people started posting messages on Twitter pleading for help. A baby was saved after the following message, along with a picture of a sleeping baby, was posted "Please help us she is a newborn" (Koren, 2017). Such posts are of great importance for response organizations as they contain real-time information shared by eyewitnesses during ongoing disasters and can save lives (Roy et al., 2021).

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

Dinani, S. and Caragea, D.
A Comparison Study for Disaster Tweet Classification Using Deep Learning Models.
DOI: 10.5220/0012129300003541
In *Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA 2023)*, pages 152-163
ISBN: 978-989-758-664-4; ISSN: 2184-285X

fected population (Alam et al., 2018).

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

Another recent deep learning model which has also shown high potential in the text classification field is CapsNet model, first introduced for image classification (Sabour et al., 2017), which has been shown to outperform baseline models such as CNN, SVM-based models, and LSTM-based models in tasks such as sentiment classification (Wu et al., 2020; Zhang et al., 2018; Zhao et al., 2018), question categorization (Zhao et al., 2018), question answering (Zhao et al., 2019a), news categorization (Zhao et al., 2018), and text classification (Chen et al., 2023). The superiority of CapsNet models has been demonstrated using various capsule architectures, including NLP-Capsule and CapsuleDAR (Zhang et al., 2018). In several studies, CapsNet models have exceeded baselines on multiple datasets and have proven their capability in transferring information from single-label to multi-label text classification (Zhao et al., 2018).

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

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

2 RELATED WORK

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

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

Most recently, several studies have explored the use of transformer-based models for disaster tweets

classification (Chanda, 2021; Ningsih and Hadiana, 2021; Zahera et al., 2019; Maharani, 2020; Naaz et al., 2021; Ray Chowdhury et al., 2020). For example, (Ma, 2019) showed that the BERT model outperformed the Bi-LSTM model but had the similar performance to customised BERT-based models that combined BERT with LSTM or CNN networks (i.e., BERT+LSTM and BERT+CNN). (Alam et al., 2020) compared three transformer-based models (BERT, DistilBERT, RoBERTa) with CNN and FastText models for both informativeness and humanitarian tasks on three datasets (one of them being the benchmark dataset, named CrisisBench). The authors showed that the transformer-based models outperformed the CNN and FastText models, while the performance of the transformer-based models was similar. Given these results of prior studies, we consider BERT to be a state-of-the-art model for disaster tweet classification, and include it as a strong model in our study.

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

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

Despite the use of CapsNets for text classification and its proven superiority over CNN and RNN models, to the best of our knowledge, there is no prior study that extensively compares CapsNets with BERT-type models on text classification. Furthermore, CapsNets have not been explored for disaster tweet classification, although they have shown promising results on disaster image classification (Dinani and Caragea, 2021). Therefore, the objective of our work was to perform a comprehensive comparison between CapsNet and BERT models on disaster tweet classification. We focused on CapsNets with both static and dynamic routings (Kim et al., 2020), among which static routing has given better results in prior works (Demotte et al., 2021; Kim et al., 2020), and addressed two specific tasks useful for disaster response: filtering tweets based on informativeness (a binary classification task) and identifying a tweet's humanitarian category (a multi-class classification task). We used three existing datasets in our experiments: CrisisLex, CrisisNLP, and CrisisBench, a benchmark dataset consolidating eight existing datasets (Alam et al., 2020). In addition to the CapsNet and BERT models, we include LSTM/Bi-LSTM models in our study given that these models have been used as strong baselines for CapsNet and BERT models in prior works (Wu et al., 2020; Zhang et al., 2018; Zhao et al., 2019a; Zhao et al., 2018).

3 METHODS

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

3.1 LSTM Networks

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

vanishing or exploding gradient problems faced by vanilla RNNs. More specifically, an LSTM unit includes a memory cell and a forget gate, together with input and output gates. The memory cell collects and remembers information, while the forget gate decides when and how much of the old information to forget. The input and output gates are in charge of controlling the amount of information that goes into and out of the cell, respectively (Hochreiter and Schmidhuber, 1997; Liu and Guo, 2019).

3.2 Bidirectional LSTM (Bi-LSTM)

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

3.3 Bidirectional Encoder Representations from Transformers (BERT)

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

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

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

3.4 Capsule Networks

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

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

A routing-by-agreement algorithm is used to determine the part-whole relationships between lower-layer capsules (parts) and higher-layer capsules (wholes). According to this algorithm, if several predictions made by active lower-layer capsules (through transformation matrices) agree, a higher-layer capsule becomes active (Sabour et al., 2017). For the first time, (Zhao et al., 2018) demonstrated the effectiveness of CapsNets for text classification. For the purpose of stabilizing the dynamic routing process and mitigating the noise caused by capsules not being effectively trained or by capsule with "back-

ground" information of the input sequence (e.g., stop words and unrelated words with respect to specific categories), they proposed three strategies to enhance the performance of dynamic routing algorithm. The first strategy was to include an extra "Orphan" category to CapsNet with the purpose of capturing the "background" knowledge. This extra category causes CapsNet to more efficiently learn the part-whole relationship. The second strategy was to use Leaky-softmax instead of standard softmax for calculating connection strength between the part-whole capsules. The third strategy was to use lower-layer capsule's probability of existence for iteratively updating the connection strength.

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

4 IMPLEMENTATION DETAILS

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

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

4.1 Datasets

The datasets used in this study include CrisisLex, CrisisNLP, and CrisisBench, the statistics of which can be found in Tables 1, 2, and 3, respectively.

Table 1: Statistics for CrisisLex dataset.

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

Table 2: Statistics for CrisisNLP dataset.

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

All three datasets have annotations for both informativeness and humanitarian category tasks. For all three datasets, the informativeness task includes two classes of Informative and Not-informative; however, for the humanitarian category task, there are 6, 10 and 11 classes in CrisisLex, CrisisNLP and CrisisBench datasets, respectively, as shown in the corresponding tables. A brief description of each dataset is given below.

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

Table 3: Statistics for CrisisBench dataset.

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

4.2 Performance Metrics

In the experiments, we compare the CapsNet models (both with static and dynamic routing algorithms, shown by CapsNet-s and CapsNet-d, respectively) with BERT models and also with LSTM and Bi-LSTM models on both binary (tweet informativeness) and multi-class (tweet humanitarian category) classification tasks. For both types of tasks, the comparison is performed using several standard evaluation metrics, specifically, Precision, Recall and F1 scores for each class and Weighted Precision, Weighted Recall and Weighted F1 scores for the overall performance of each model.

4.3 Experimental Setup

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

The hyperparameters employed in the experiments were obtained through fine-tuning on the respective development (dev) datasets. Specifically, for the LSTM and Bi-LSTM experiments, the following hyperparameters were utilized: hidden size of 256, maximum sequence length of 60, learning rate of 5e-4, drop ratio of 0.1, weight decay of 0 (i.e., no L2 regularization), a batch size of 40, maximum number of epochs of 10.

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

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

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

5 EXPERIMENTAL RESULTS AND DISCUSSIONS

The experimental results of crisis tweet classification for informativeness category task of all three datasets are presented in Table 4 and for humanitarian category task of CrisisLex, CrisisNLP, and CrisisBench are presented in Tables 5, 6, and 7, respectively. The tables depict the performance of the models for each class (namely, Precision, Recall, and F1-score) in the datasets, along with the overall performance of the models in the "Overall" row, which shows the Weighted Precision, Weighted Recall, and Weighted F1-score across all classes of each dataset. We discuss the answers to our questions using these results in the following paragraphs.

Our first research question was to compare CapsNet models to state-of-the-art BERT models and also to LSTM/Bi-LSTM models trained to classify crisis-related tweets in terms of informativeness (binary classification) and humanitarian category (multi-class classification) using three datasets, specifically CrisisLex, CrisisNLP, and CrisisBench. For each task, we initially analyze the overall performance of the models and then examine the results for each class in the dataset. As can be seen in Tables 4, 5, 6, and 7 for both tasks, BERT models are the best models in terms of all the evaluation metrics considered, when compared to LSTM, Bi-LSTM, CapsNet-s and Capsnet-d models.

For the tweet informativeness task, the F1-scores of BERT models are 94.571, 85.867, and 87.975 for CrisisLex, CrisisNLP, and CrisisBench datasets, respectively. These scores are higher than those of the other four models for the corresponding datasets. For this classification task and for the two datasets of CrisisLex, CrisisNLP, the order of other models, from best to worst, is Bi-LSTM, LSTM, CapsNet-d and CapsNet-s (with F1-scores of 94.477, 94.443, 94.110, 93.922, respectively, for CrisisLex dataset and 83.786, 83.500, 82.986, 82.654, respectively, for

Table 4: Experiment results for tweet informativeness task for all three datasets, for each class and the overall weighted metrics, with the best value of each metric displayed in bold within each row.

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

Table 5: Experiment results for tweet humanitarian task for the CrisisLex datasets, for each class and the overall weighted metrics, with the best value of each metric displayed in bold within each row.

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

Table 6: Experiment results for tweet humanitarian task for the CrisisNLP datasets, for each class and the overall weighted metrics, with the best value of each metric displayed in bold within each row.

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

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

For the crisis tweet humanitarian category, the F1-scores of the BERT models are 93.008, 81.300, and 86.708 for CrisisLex, CrisisNLP, and CrisisBench datasets, respectively. As for the crisis tweet informativeness tasks, these scores are higher than those of the other four models for the corresponding datasets. For this classification task and for the dataset of CrisisLex, the order of other models, is LSTM, Bi-LSTM, CapsNet-s and CapsNet-d (with F1-scores of 92.145, 91.850, 91.609, and 91.539, respectively), while for the other two datasets, the order is Bi-LSTM, LSTM, CapsNet-s and CapsNet-d (with F1-scores of 77.089, 76.784, 75.302, 73.991, respectively, for CrisisNLP dataset and 83.988, 83.528, 81.778, 81.757, respectively, for CrisisBench dataset). BERT outperforms other models in terms of F1-scores on all classes in CrisisNLP and CrisisBench datasets. In CrisisLex, LSTM leads only in the "Not humanitarian" class with an F1-score of 97.731, while BERT is close behind with a score of 97.702. For the rest of the classes in CrisisLex, BERT still outperforms other models. BERT's classifications of instances in different classes can be better understood through the confusion matrices (CM) in Figures 1, 2, and 3 for the three datasets. As can be seen in these

Table 7: Experiment results for tweet humanitarian task for the CrisisNLP datasets, for each class and the overall weighted metrics, with the best value of each metric displayed in bold within each row.

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

Figures, for the CrisisLex dataset, the model seems to have difficulties in correctly identifying instances that relate to "Infrastructure and utilities damage", with the accuracy of 0.59. For CrisisNLP, most of the instances of "Requests or needs" and "Response efforts" classes (with the lowest accuracy of 0.14 and 0.37, respectively) have been mis-classified as "Not humanitarian". Similarly, in CrisisBench, most of the instances of "Response efforts" class (with the lowest accuracy of 0.30) have been mis-classified as "Not humanitarian". Based on these results, we can conclude that BERT is the best model for classifying crisis-related tweet datasets, while CapsNet models (with both static and dynamic routing) have the worst performance. Therefore, the properties of CapsNet do not seem to benefit much the task of classifying crisis-related tweet datasets.

Our second research question was to compare two routing algorithms used in CapsNet models, mainly static and dynamic routings, for classifying crisis-related tweets in terms of informativeness and humanitarian category using the three datasets considered in our study. For the crisis tweet informativeness task and for two datasets (CrisisLex, CrisisNLP), the performance of CapsNet-d is better than that of CapsNet-s (with F1-scores of 94.110, 93.922, respectively, for the CrisisLex dataset, and 82.986, and 82.654, respectively, for the CrisisNLP dataset). For the CrisisBench dataset, the F1-score of CapsNet-s (86.358) is higher than that of CapsNet-d (86.320). However, for the tweet humanitarian category task, for all three datasets, CapsNet-s outperforms CapsNet-d, with F1-scores of 91.609 versus 91.539, respectively, for CrisisLex, 75.302 versus 73.991, respectively, for CrisisNLP dataset, and 81.778 versus 81.757, respec-

tively, for the CrisisBench dataset. However, in all cases, the F1-scores of the two routing algorithms are pretty close to each other. Based on the results, we can conclude that when classifying crisis-related tweets both in terms of informativeness (binary classification) and humanitarian category (multi-class classification), both static routing and dynamic routing algorithms have similar performance

Overall, based on the obtained results, BERT emerges as a stronger candidate compared to CapsNet for disaster tweet classifications. The superior performance of BERT can be attributed to several key factors. Firstly, BERT's pre-training on extensive data enables it to acquire a deep understanding of complex patterns and contextual information. Additionally, its utilization of the attention mechanisms and transformer architecture empowers BERT to capture long-range dependencies in language, resulting in a better comprehension and generation of natural language. While CapsNet shows promise in specific image-related tasks, thanks to its equivariant property that preserves spatial relationships between components, it has not attained the same level of success as BERT in the field of natural language processing.

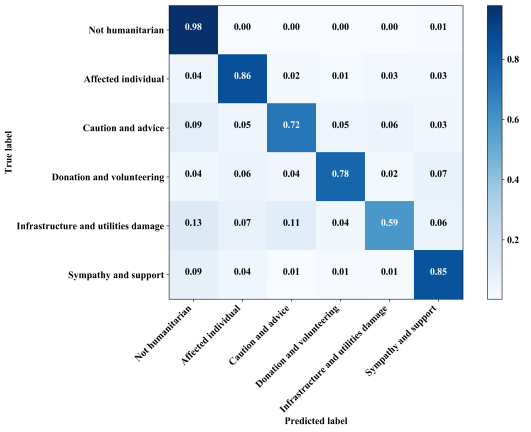

Figure 1: BERT's CM for humanitarian task on CrisisLex.

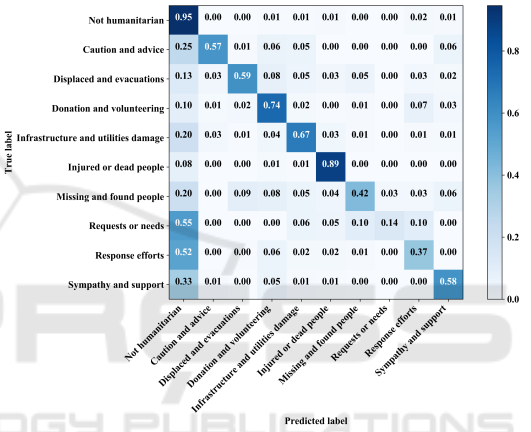

Figure 2: BERT's CM for humanitarian task on CrisisNLP.

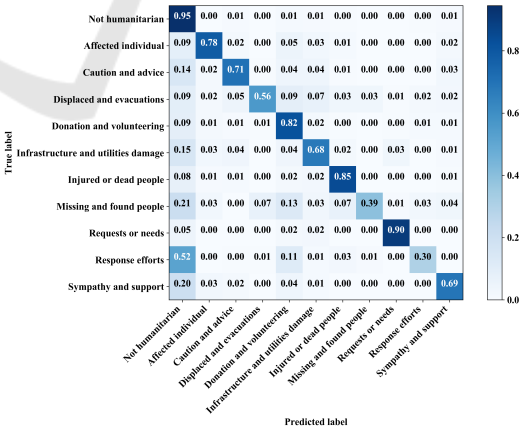

Figure 3: BERT's CM for humanitarian task on CrisisBench.

6 CONCLUSIONS AND FUTURE WORK

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

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

ACKNOWLEDGEMENTS

We 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.

REFERENCES

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

Ahmed, A. (2011). Use of social media in disaster management.

Alam, F., Alam, T., Hasan, M. A., Hasnat, A., Imran, M., and Ofli, F. (2023). Medic: a multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, 35(3):2609–2632.

Alam, F., Ofli, F., Imran, M., and Aupetit, M. (2018). A twitter tale of three hurricanes: Harvey, Irma, and Maria. *arXiv preprint arXiv:1805.05144*.

Alam, F., Sajjad, H., Imran, M., and Ofli, F. (2020). Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. *arXiv preprint arXiv:2004.06774*.

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

Chanda, A. K. (2021). Efficacy of bert embeddings on predicting disaster from twitter data. *arXiv preprint arXiv:2108.10698*.

Chen, Z., Li, S., Ye, L., and Zhang, H. (2023). Multi-label classification of legal text based on label embedding and capsule network. *Applied Intelligence*, 53(6):6873–6886.

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

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

Dinani, S. T. and Caragea, D. (2021). Disaster image classification using capsule networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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

Hinton, G. E., Sabour, S., and Frosst, N. (2018). Matrix capsules with em routing. In *International conference on learning representations*.

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

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2018). Processing social media messages in mass emergency: Survey summary. In *Companion Proceedings of the The Web Conference 2018*, pages 507–511.

Imran, M., Mitra, P., and Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.

Imran, M., Ofli, F., Caragea, D., and Torralba, A. (2020). Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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

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

Khajwal, A. B., Cheng, C.-S., and Noshadravan, A. (2023). Post-disaster damage classification based on deep multi-view image fusion. *Computer-Aided Civil and Infrastructure Engineering*, 38(4):528–544.

Khikmatullaev, A. (2019). *Capsule Neural Networks for Text Classification*. Universit¨at Bonn.

Kim, J., Jang, S., Park, E., and Choi, S. (2020). Text classification using capsules. *Neurocomputing*, 376:214–221.

Koren, M. (2017). Using twitter to save a newborn from a flood. The Atlantic.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4):150.

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

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

Liu, G. and Guo, J. (2019). Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, G., Gan, J., Yin, J., Luo, Z., Li, B., and Zhao, X. (2020). Multi-task learning using a hybrid representation for

text classification. *Neural Computing and Applications*, 32(11):6467–6480.

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

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

Mazloom, R., Li, H., Caragea, D., Caragea, C., and Imran, M. (2019). A hybrid domain adaptation approach for identifying crisis-relevant tweets. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 11(2):1–19.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Naaz, S., Abedin, Z. U., and Rizvi, D. R. (2021). Sequence classification of tweets with transfer learning via bert in the field of disaster management. *EAI Endorsed Transactions on Scalable Information Systems*, 8(31):e8.

Neppalli, V. K., Caragea, C., and Caragea, D. (2018). Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In *Proceedings of the 15th Annual Conference for Information Systems for Crisis Response and Management (ISCRAM)*.

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

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Eighth international AAAI conference on weblogs and social media*.

Patrick, M. K., Adekoya, A. F., Mighty, A. A., and Edward, B. Y. (2022). Capsule networks–a survey. *Journal of King Saud University - computer and information sciences*, 34(1):1295–1310.

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

Roy, P. K., Kumar, A., Singh, J. P., Dwivedi, Y. K., Rana, N. P., and Raman, R. (2021). Disaster related social media content processing for sustainable cities. *Sustainable Cities and Society*, 75:103363.

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in neural information processing systems*, 30.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sathishkumar, V. E., Cho, J., Subramanian, M., and Naren, O. S. (2023). Forest fire and smoke detection using deep learning-based learning without forgetting. *Fire Ecology*, 19(1):1–17.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Villegas, C., Martinez, M., and Krause, M. (2018). Lessons from harvey: Crisis informatics for urban resilience. *Rice University Kinder Institute for Urban Research*.

Wu, Y., Li, J., Wu, J., and Chang, J. (2020). Siamese capsule networks with global and local features for text classification. *Neurocomputing*, 390:88–98.

Yang, M., Zhao, W., Chen, L., Qu, Q., Zhao, Z., and Shen, Y. (2019). Investigating the transferring capability of capsule networks for text classification. *Neural Networks*, 118:247–261.

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

Zahera, H. M., Elgendy, I. A., Jalota, R., and Sherif, M. A. (2019). Fine-tuned bert model for multi-label tweets classification. In *TREC*, pages 1–7.

Zhang, B., Xu, X., Yang, M., Chen, X., and Ye, Y. (2018). Cross-domain sentiment classification by capsule network with semantic rules. *IEEE Access*, 6:58284–58294.

Zhao, W., Peng, H., Eger, S., Cambria, E., and Yang, M. (2019a). Towards scalable and reliable capsule networks for challenging nlp applications. *arXiv preprint arXiv:1906.02829*.

Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S., and Zhao, Z. (2018). Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538*.

Zhao, Y., Shen, Y., and Yao, J. (2019b). Recurrent neural network for text classification with hierarchical multiscale dense connections. In *IJCAI*, pages 5450–5456.