

Disaster Classification and Progression Detection from Social Media Data Using Deep Learning

Md Saimul Haque Shanto¹, Suzad Mohammad^{1,2}, Ibna Masud Rana Parvej¹,
Manzur Kabir Ayon¹, and Sadia Sharmin¹

¹ Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Bangladesh
{0424052002, 0423052030, 1024052034, 0423052097}@grad.cse.buet.ac.bd,
sadiasharmin@cse.buet.ac.bd

² Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Bangladesh

Abstract. In today's world, social media platforms such as Twitter, Facebook have become vital sources of real-time information during disasters, providing us useful data that can help save lives and reduce damage if properly analyzed. Our study examines how to utilize this data to categorize disaster types and monitor their progression stages. We have worked with a combined dataset (CrisisLex T6, CrisisLex T26, and CrisisNLP), which includes over 42,000 tweets covering 13 types of natural and human-made disasters which are split into three stages: early signs, peak impact, and recovery. We have prepared the data by cleaning the raw text data (tweets or posts) removing URLs, mentions, and hashtags, and then balanced the dataset using oversampling and undersampling to handle uneven class distributions. Our approach uses word embeddings as the foundation, feeding them into a Bidirectional Long Short-Term Memory (Bidirectional LSTM) model to capture the context of tweets from both directions. We dropped out some layers to avoid overfitting and finished with dense layers and also a softmax function for making the best predictions. We ran the experiments on Google Colab with 30 epochs, a learning rate of 0.001, a batch size of 64, and the ADAM optimizer. We achieved impressive results with 99% accuracy, 99% precision, 99% recall, and 99% F1-score for classifying disaster types, and 96% accuracy, 95% precision, 87% recall, and 91% F1-score for tracking progression stages. These outcomes beat the previous papers' results by increasing up to 5% in F1-score, showing that our method can significantly improve the detection of disasters.

Keywords: CrisisLex T6 · CrisisLex T26 · CrisisNLP · Oversampling · Undersampling · Bidirectional Long Short-Term Memory (Bidirectional LSTM) · ADAM optimizer.

1 Introduction

Social media sites (Facebook, Twitter, Instagram, LinkedIn etc.) have emerged as an effective mechanism for spreading information in real time, specially dur-

ing calamities in the internet era [25, 12, 19]. They tend to have more eyewitness tweets, accounts, updates, and calls for assistance than traditional news sources [13]. Previous studies suggest that these sites produce a massive amount of data, with millions of messages for each major event [24, 18]. These data provide insights into crises, which can range from man-made catastrophes to natural disasters. Previous researches primarily concern sentiment analysis [26, 21] and keyword-based filtering [6, 1]. Though complex approaches, such as neural networks [8], in classifying disaster-related content had been challenging to utilize, deep learning breakthroughs have since made it possible. Data collections like CrisisLex and CrisisNLP formed the foundation on which disaster response automated systems were built, serving a fundamental role in training the models.

In spite of all these developments, there are still some obstacles in the way of efficient use of social media data for disaster relief [15]. It is not possible for manual analysis about the nature of tweets filled with emotions, slang terms, abbreviations, and irrelevant information [7]. Unreliability of the traditional means, such as rule-based classifiers or basic machine learning algorithms, to detect types of catastrophes and monitor their phases from their first occurrence to the point of greatest damage and recovery, deprives response and resource provision [4]. For catastrophic incidents this becomes more important because detection early can save lives and reduce economic loss [9]. We are driven by the challenge of bridging the gap with deep learning [10] to build a more accurate, automated system that not only identifies disasters but also tracks their evolution, eventually helping governments, non-governmental agencies, and first responders make defensible decisions in extenuating situations.

We contributed to this project by developing a comprehensive deep learning pipeline designed for the categorization of disasters and the detection of progression using social media data. With preprocessing steps including text cleaning (removing URLs, mentions, and hashtags) and class balancing through oversampling and undersampling, we assembled a combined dataset of more than 42,000 tweets from CrisisLex [17] T6, T26, and CrisisNLP [3], covering 13 disaster categories and three progression stages. We use Bidirectional Long Short-Term Memory (Bidirectional LSTM) network at the center trained for 30 epochs with patience 5 using ADAM optimizer with word embeddings, dropout layers, and softmax activation added to it. The method provided outstanding performance, bettering past records by up to 5% in F1-score and reaching 99% accuracy in disaster type classification and 96% accuracy in progression detection. By using comparison studies and confusion matrices to show real implementations, our work offers the potential for real-time, scalable catastrophe monitoring systems.

An overview of our proposed approach is depicted in figure 1

2 Literature Review

Social media for disaster management has progressed a long way, with recent studies emphasizing deep learning techniques for the management of vast, unstructured data from Twitter. Mukhtiar et al. proposed a relevance classifica-

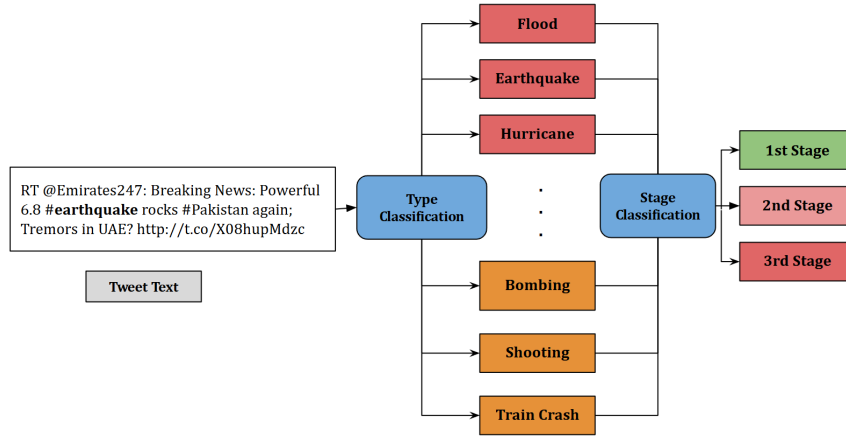


Fig. 1: Proposed Workflow

tion model for Twitter tweets on floods utilizing a number of transformer models—BERT, RoBERTa, and XLNet—fine-tuned individually and ensembled using late fusion method to combine the prediction scores, with text preprocessing and hyperparameters optimized such as learning rates of 0.001 for BERT and RoBERTa, and 0.002 for XLNet, with batch sizes between 8 and 32 over 3 to 10 epochs and with high accuracy of 0.93 and F1-scores up to 0.94 for BERT, although limited to flood events and not generalized to multi-disasters [16]. Kabir et al. designed a multi-headed deep neural network with Attention-based Bidirectional LSTM and CNN for disaster tweet classification and scheduling of rescue, using data from Hurricanes Harvey and Irma collected by Twitter’s Stream API to classify the tweets into such requirements as rescue, water, or medical assistance, using GloVe embeddings and basing against benchmarks like CrisisNLP and CrisisLex, achieving an F1-score of 0.93 with an adaptive schedule strategy that kept the priorities in balance under resource constraints, but hindered by limited labeled data for low-frequency classes like "injured" and absence of location mentions in the majority of tweets [14]. Alam et al. introduced the HumAID dataset, comprising approximately 77,000 human-annotated tweets from 19 disaster events focused on natural calamities, providing benchmarks for classical models (e.g., SVM, RF) and deep learning approaches (e.g., fastText, BERT), addressing class imbalance across 11 categories with English-only annotations, and reporting weighted F1-scores averaging 0.78 to 0.87, while noting exclusions of human-made disasters and potential biases from imbalanced distributions [2]. Building on transfer learning, Saleem and Mehrotra presented a context-aware framework using DistilBERT with multistage fine-tuning and feed-forward neural networks to detect informative social media content during disasters, applied to the CrisisMMD dataset of about 8,600 tweets from seven events like hurricanes and earthquakes, tackling class imbalance via down-sampling and achieving F1-

scores of 0.82 to 0.94, outperforming BERT and RoBERTa baselines, though its binary classification (informative vs. non-informative) and reliance on static datasets limit its scope for progression tracking or real-time deployment [22]. Dharrao et al. advanced disaster tweet classification with a gradient-based optimized CNN incorporating BERT embeddings and RMSprop optimization, tested on Kaggle’s “Disaster Tweets” dataset of around 10,000 binary-labeled entries (disaster vs. non-disaster), involving text preprocessing and model training to result in an accuracy of 0.83 and F1-score of 0.80, surpassing logistic regression and decision trees, while its limitations include a narrow focus on binary tasks without exploring multi-class scenarios or real-time integration [5]. Collectively, these works underscore the efficacy of transformers and hybrid neural architectures in handling noisy social media data, yet reveal gaps in multi-disaster progression detection, dataset diversity, and practical scalability that our study aims to address.

3 Methodology

3.1 Dataset Description

For our experiment, we use a combined dataset by pooling together three broadly accepted benchmark sets in the field of crisis informatics literature: CrisisLex T6 [17], CrisisLex T26 [18], and CrisisNLP [13]. These datasets include real tweets from natural and man-made disasters, like earthquakes, floods, hurricanes, wildfires, explosions, shootings and so on.

The CrisisLexT6 contains tweets related to six natural disasters, primarily earthquakes, hurricanes, and floods. CrisisLex T26 expands the coverage scope to up to 26 disaster events, both natural disasters such as wildfires and typhoons and human-made disasters such as train collisions and bombings. There are seven disaster scenarios (e.g., earthquake, flood, typhoon, and volcanic eruption) tweets annotated for disaster-related information tasks available in CrisisNLP.

Subsequently, through intense preprocessing and merge operations, the final dataset contains 42,953 annotated tweets spanning over 13 disaster classes. To enable the analysis of the evolution of disasters over time, the dataset was annotated over three progression phases: (i) first phase (preliminary warning), (ii) stage two (intensification and extensive impacts), and (iii) stage three (recovery and aftermath). The entire dataset was split into training (70%), validation (15%), and test (15%) sets in a single stratified split, ensuring that the proportional representation of each disaster class was preserved across all three partitions. Figure 2 and Figure 3 shows the class-wise distribution of tweets across training, validation, and test sets.

Besides ensuring balanced coverage for man-made and natural disasters, the aggregate dataset ensured stage-wise annotation and hence became especially appropriate for multi-class classification and analysis of development tasks under the applications of disaster informatics.

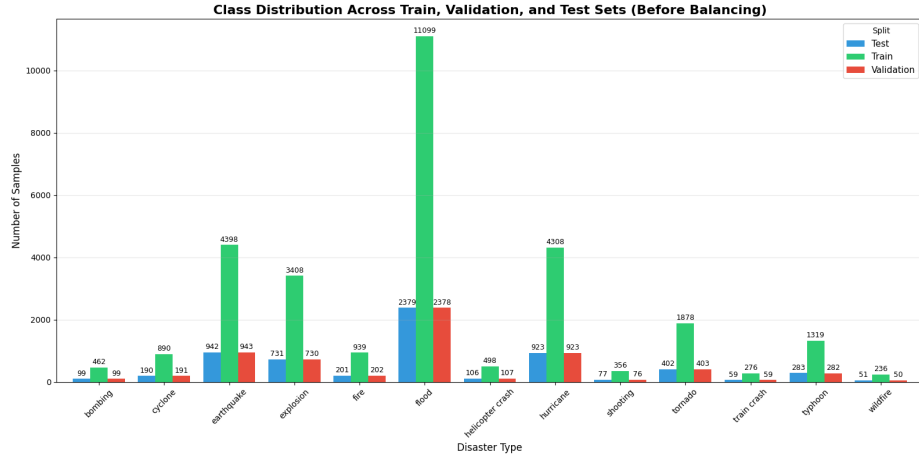


Fig. 2: Class-wise distribution of tweets across training, validation, and test sets for disaster type classification

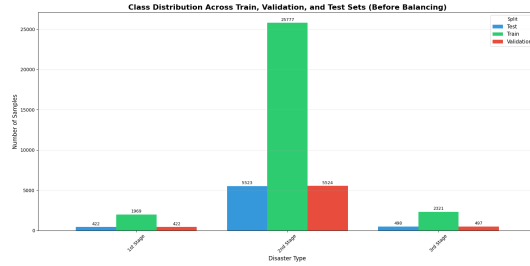


Fig. 3: Class-wise distribution of tweets across training, validation, and test sets for disaster stage classification

3.2 Data Annotation

The progression-stage labels (1st Stage, 2nd Stage, and 3rd Stage) were assigned through a rigorous, structured manual annotation process. After merging the CrisisLex T6, CrisisLex T26, and CrisisNLP datasets to create our combined corpus of 42,953 tweets, five independent annotators were involved in the labeling task. Before beginning annotation, we prepared detailed guidelines that clearly defined the three stages based on the temporal progression of disaster events:

- 1st Stage (Early / Pre-Disaster / Onset Stage): Posts indicating early warnings, onset signals, forecasts, or the initial phase of a disaster with limited or no confirmed impact. Key linguistic markers included keywords such as "alert," "warning," "forecast," "approaching," and "predicted."
- 2nd Stage (Active / During-Disaster / Impact Stage): Posts describing the active impact phase, including ongoing damage, casualties, urgent needs,

- and real-time disruption. This stage typically contains urgency markers like "happening now," "emergency," "live," "damage," and "casualties."
- 3rd Stage (Post-Disaster / Recovery / Aftermath Stage): Posts related to the aftermath, recovery efforts, damage assessment, relief distribution, and long-term consequences. Recovery-related terms such as "recovery," "reconstruction," "relief," "donation," and "rebuilding" were common in this stage.

Each tweet was independently labeled by all five annotators. We employed majority voting to resolve disagreements, and in cases where no clear majority emerged, the final label was decided through a structured discussion among the annotators to reach consensus. Inter-annotator agreement was measured using Fleiss' Kappa, which yielded a score of 0.81, indicating substantial to almost perfect agreement across all annotators.

3.3 Data Preprocessing

Social media (like Twitter) data is noisy and unstructured by its own definitions; a comprehensive preprocessing pipeline was followed before the model was trained. The raw text went through cleaning by the removal of URLs, user mentions, and unnecessary special characters, with only important keywords from hashtags preserved. The text was converted into lowercase for uniformity and repeated characters, emojis, and unnecessary punctuation were removed for the least amount of distortion. The tweets were tokenized as single words and lemmatized for normalizing the word inflections into their base form (flooding $\rightarrow$ flood). Stopwords such as the ordinary function words (the, is, at) were removed for the purpose of noise reduction and informative token focus. In addressing the class imbalance present within the dataset—inherent because the categories for disasters were not evenly distributed—methods such as class weighting, down-sampling, and upsampling were employed at training such that the minority classes were handled without losing considerable information.

3.4 Model Architecture

In this study, we propose a Bidirectional Long Short-Term Memory (Bidirectional LSTM)-based architecture for disaster types and stage classification. The framework consists of several sequential stages, as shown in Figure 4. First, we collected raw tweets and then preprocessed the data by removing noise like URLs, mentions, hashtags, and special characters along with tokenization and padding to make sure that the length of the inputs was equal. To deal with the problem of imbalance in the dataset, we applied a hybrid resampling technique combining oversampling [11] of minority classes with undersampling of the majority class.

Next, the preprocessed tokens are transformed into dense vectors through the word embedding layer, where the word's semantic content is embedded into them. They undergo a Bidirectional LSTM (Bidirectional LSTM) [23], which captures contextual dependencies both forward and backward along the tweet.

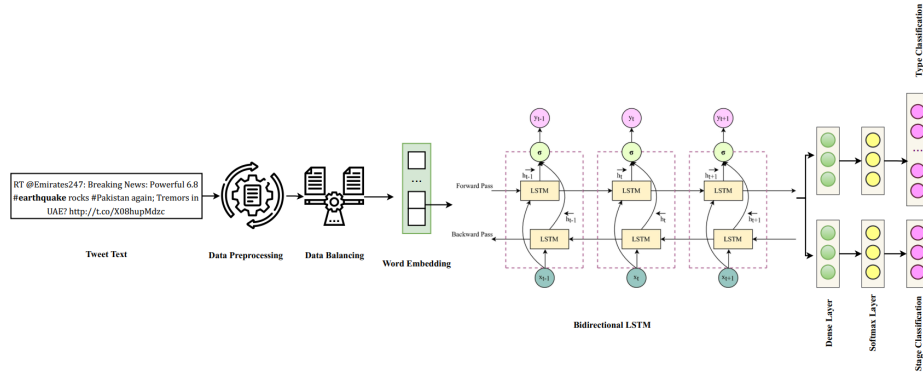


Fig. 4: An overview of the applied methodology approach.

Both the forward and backward hidden states are combined together, forming an enriched semantic representation.

The output of the Bidirectional LSTM is sent to a Multilayer Perceptron (MLP) which contains a dropout layer to reduce overfitting, then a fully connected dense layer for transforming features, and a softmax layer for generating probability distributions over target labels. The system is capable of carrying out two classification tasks: (i) type classification (identification of the disaster type, such as earthquake, flood, or hurricane), and (ii) stage classification (detection of the stage of the disaster, such as early warning, ongoing, or aftermath). It enables the model to jointly extract the syntactic as well as the semantic content effectively from noisy social media texts for the task of disaster event classification.

An overview of the entire model is depicted in Figure 4.

4 Results and Discussions

4.1 Experimental Setup

We conducted the experiments using the Google Colab Platform and trained the models on T4 GPU. We trained the Bidirectional LSTM–MLP model using different numbers of steps and applied early stopping to mitigate the overfitting problem. Table 1 shows the experimental setup used for model training across Disaster types as well as stages. The configurations include the maximum number of epochs, learning rate, batch size, and optimizer settings for each classification.

4.2 Disaster Type Classification

The first experiment was to classify the disaster types across 13 categories, including earthquakes, hurricanes, floods, wildfires, shootings, explosions and so on. The confusion matrix (Figure 5) highlights its comparable performance, with

Table 1: Experimental setup for disaster type and stage classification tasks.

Parameter	Training Parameters
Max Epoch	30
Learning Rate	0.001
Batch Size	64
Vocabulary size	15,000 (max_ words)
Sequence length	120 (max_ len)
Embedding Source	Trainable (learned from scratch)
Embedding dimension	128
Bidirectional LSTM Layer 1	128 units (256 output)
Bidirectional LSTM Layer 2	64 units (128 output)
Dense Layer 1	128 units (ReLU)
Dense Layer 2	64 units (ReLU)
Output Layer	Softmax
Dropout (LSTM Layer 1)	0.3 (input), 0.3 (recurrent)
Dropout (LSTM Layer 2)	0.3 (input), 0.3 (recurrent)
Dropout (after LSTM Layer 1)	0.4
Dropout (after LSTM Layer 2)	0.4
Dropout (after Dense Layer 1)	0.5
Dropout (after Dense Layer 2)	0.3
Random seed	42
Optimizer	Adam

minimal misclassifications. Disasters like explosions, earthquakes and hurricanes reflect a strong predictive score, which explains the robustness of the proposed methodology. However, the other classes, like tornado, flood, are sometimes confused with each other because of the semantic similarities of these categories. Since both of them are water-related natural disasters and often share common textual descriptors such as “water”, it sometimes becomes difficult to distinguish between them.

Figure 5’s heatmap report also confirms the performance of each disaster class with different evaluation metrics such as precision, recall, and F1-scores. In all categories, the model achieves greater than 0.96 values consistently, whereas some classes, such as earthquake, explosion, and helicopter crash, result in perfect or near-perfect precision and recall. The overall accuracy is 0.99, and macro-averaged F1-score is 0.99, confirming well-balanced performance across all kinds of natural and human-made disasters.

To assess the robustness and reproducibility of the findings obtained from our model, the experiment was conducted multiple times which resulted in four distinct trials. Table 3 illustrates the averages and standard deviations of the metrics evaluated during each trial.

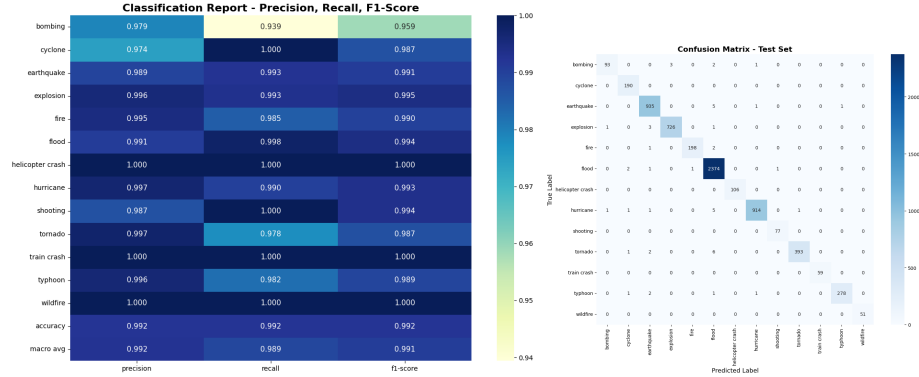


Fig. 5: Experimental result of Disaster Type Classifier

Table 2: Performance of disaster type classification.

Metric	Score
Accuracy	0.99
Precision	0.99
Recall	0.99
F1 Score	0.99

4.3 Disaster Stage Classification

Figure 6 shows the confusion matrix of the classification of the disaster stage task. The model exhibits good classification ability for all three stages. In particular, the 2nd Stage class achieves the highest recognition rate, where it correctly classifies up to 5,478 samples and only slightly misclassifies (25 as 1st Stage and 20 as 3rd Stage). In the 1st Stage, it correctly identified 334 samples with fair confusion with the 2nd Stage (86 cases) and very few misclassified as the 3rd Stage (2 cases). In the 3rd Stage, it correctly identified 413 instances, though it misclassified up to 85 samples as 2nd Stage, indicating partial overlap of feature representation for the two stages.

Figure 6 provides the classification report as a heatmap where precision, recall, and F1-score for every stage of the disaster are indicated. The 2nd Stage exhibits the best-performing reliability with precision (0.97), recall (0.99), and F1-score (0.98). The 1st Stage reveals balanced performance with 0.93 precision, 0.79 recall, and 0.86 F1-score, demonstrating good classification ability. The 3rd Stage, although still fairly performing, exhibits slightly lower results, specially for recall (0.83), which implies some cases were mislabeled into nearby stages.

These results from the confusion matrix and heatmap confirm the robustness of the proposed model in distinguishing disaster stages, with particularly high confidence in detecting the critical 2nd Stage, which is important as it provides valuable insights into the ongoing disaster.

Table 3: Performance metrics across multiple runs for type classification.

Run	Precision	Recall	F1-score	Accuracy
1	0.99	0.99	0.99	0.99
2	0.98	0.97	0.98	0.98
3	0.99	0.98	0.98	0.99
4	0.98	0.98	0.98	0.98
Mean $\pm$ Std	0.985 $\pm$ 0.006	0.980 $\pm$ 0.008	0.982 $\pm$ 0.005	0.985 $\pm$ 0.006

Table 4: Performance of disaster stage classification.

Metric	Score
Accuracy	0.96
Precision	0.95
Recall	0.87
F1 Score	0.91

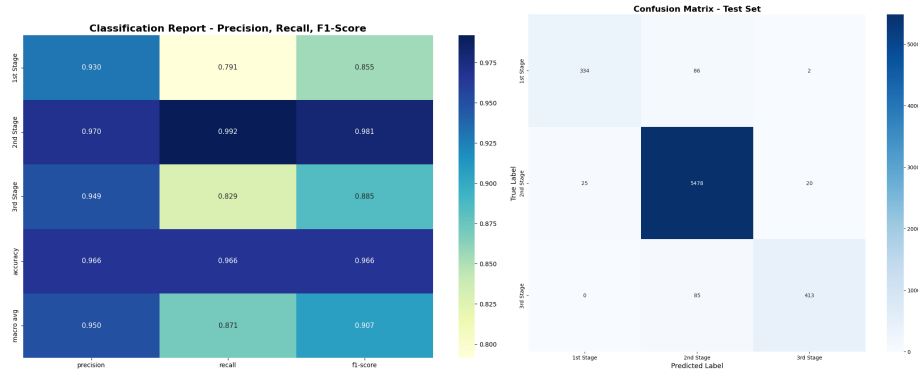


Fig. 6: Experimental result of Disaster Stage Classifier

To assess the robustness and reproducibility of the findings obtained from our model, the experiment was conducted multiple times which resulted in four distinct trials. Table 5 illustrates the averages and standard deviations of the metrics evaluated during each trial.

4.4 Explainability with LIME

In order to achieve transparency of our model, we applied Local Interpretable Model-agnostic Explanations (LIME) [20]. In particular, LIME allowed us to analyze what words from the tweets impacted the decision of the model.

Figure 7 shows the LIME explanation for a tweet that belongs to class "flood." The chart provides a visual representation of the top words used in the prediction process. We may observe that words "flood," "storms," "australia's," and "hoisted" were recognized with a relatively high score of importance, and

Table 5: Performance metrics across multiple runs for stage classification.

Run	Precision	Recall	F1-score	Accuracy
1	0.95	0.87	0.91	0.97
2	0.94	0.84	0.88	0.96
3	0.95	0.82	0.87	0.96
4	0.96	0.82	0.88	0.96
Mean $\pm$ Std	0.950 $\pm$ 0.008	0.838 $\pm$ 0.024	0.885 $\pm$ 0.017	0.963 $\pm$ 0.005

thus the model was able to use disaster-related keywords. At the same time, neutral words such as "amazing" and "bag" made no contribution to the classification or even had negative weights.

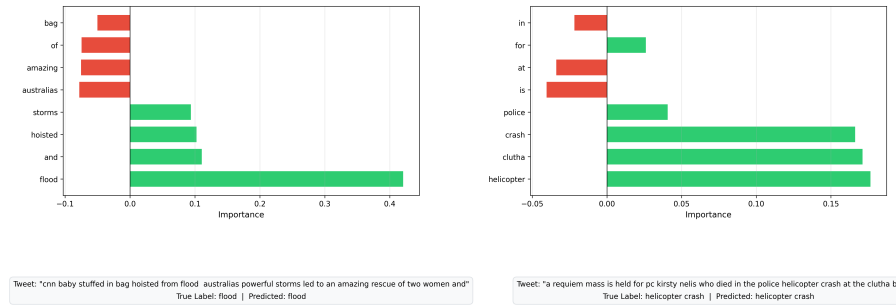


Fig. 7: XAI implemented on Tweet.

Thus, we can be confident that our model is not trained to recognize any random associations. For instance, in earthquake-related tweets, the words "tremor" and "magnitude" have high scores of importance, whereas "blast" and "casualties" were considered important for explosion-related tweets.

At the same time, LIME helps us detect if there is a confusion of the model regarding some similar words across different disasters.

4.5 Comparative Analysis

Mukhtiar et al. [16] targeted the problem of filtering disaster-related information from social media, i.e., tweets about floods. Their approach employed transformer-based models like BERT, RoBERTa, and XLNet for relevance classification. Even though their approach possessed an excellent F1-score value of around 0.88, its real-world applicability was constrained because the dataset was restricted to flood-related events alone. Such narrow restriction limits the generality of the model across more universal disaster scenarios.

Kabir et al. [14] presented a deep learning architecture for disaster tweet classification and rescue planning. They used an attention-based Bidirectional

LSTM and CNN layers to create a multi-headed architecture that was capable of extracting sequential as well as spatial features of the tweets. In another effort, they also contributed greatly to the research community by developing HumAID, a large-scale human-annotated dataset of almost 77,000 tweets in 11 categories of natural disasters. Despite these advancements, the dataset contained some limitations, such as category imbalance, lack of human-induced disasters, and restriction to English tweets. Despite these imperfections, the model performed well, with a reported F1-score of approximately 0.93.

Alam et al. [2] concentrated on enhancing disaster tweet classification by benchmarking their new HumAID dataset published shortly before. They proposed a model that involves combining DistilBERT embeddings and a feedforward neural network, which they trained through a multi-step process. Their dataset, CrisisMMD, contained approximately 8,600 tweets from seven different disaster events. The study focused mainly on binary classification—informative from non-informative tweets—rather than trying to handle multi-class disaster types. Therefore, the model's performance was moderate at an F1-score of around 0.78, observing the difficulty encountered in dealing with disaster tweet classification using sparse and unbalanced data.

Saleem et al. [22] explored transfer learning to improve disaster tweet classification. The authors proposed a hybrid approach making use of CNN and BERT-based embeddings, allowing the model to learn contextual as well as local semantic features. They carried out their experiments on the Kaggle "Disaster Tweets" dataset containing about 10,000 disaster or not labeled tweets. The model did better compared to using traditional machine learning approaches but within a small range of binary classification without giving any insights into disaster development or multi-class classification. The results they presented were an accuracy of about 0.83 and an F1-score of 0.80, which was encouraging but also indicative of the nature of the dataset.

Dharrao et al. [5] fine-tuned Convolutional Neural Network (CNN) models for disaster tweet classification using gradient-based optimization methods and BERT embeddings. Their model was also evaluated using the Kaggle disaster tweets dataset, as was Saleem and Mehrotra's. While their approach worked decently with an F1-score of around 0.80, it suffered from the small size and narrow diversity of the dataset. The use of one dataset suggested the model's performance on handling diverse disaster events was not tested. Despite these limitations, their work contributed in the sense that they illustrated how CNNs can be enhanced for disaster-related text when combined with state-of-the-art embeddings.

Our Bidirectional LSTM-MLP-based model outperformed previous work with an accuracy of 0.99 and an F1-score of 0.99 for Classification of Disaster Types and obtained 0.89 F1-score and 0.96 accuracy for Classification of Disaster Stages.

Table 6: Comparison of performance metrics between the proposed method and related works.

Previous Works	Method	Precision	Recall	F1-score	Accuracy
Alam et al.	RoBERTa	–	–	0.78	–
Kabir et al.	attention-based Bidirectional LSTM + CNN	0.94	0.94	0.93	0.94
Dharrao et al.	BERT-CNN + RMSProp	–	–	0.80	0.83
Mukhtiar et al.	RoBERTa + XLNet (Late Fusion)	0.90	0.86	0.88	–
Proposed Method	Proposed Model	0.99	0.99	0.99	0.99

5 Conclusion and Future Direction

In this paper, we utilized social media (Twitter) data for categorizing both disaster types and its progression (warning, impacts, and recovery). We proposed a deep learning-based framework combining Bidirectional LSTM and MLP, which achieved 0.99 accuracy in disaster type classification and 0.96 accuracy in progression detection, demonstrating the effectiveness of social media streams for real-time disaster monitoring. Though previous researchers’ work on natural disaster classification, there has been no work on human-made disasters and the progression of disasters. We have also shown comparative analysis with previous works, and our works done better in all metrics.

While the results are promising, there are some interesting avenues for future work. A straightforward extension is incorporating multilingual datasets, such that the system can operate within overseas disaster scenarios beyond the English-speaking emphasis of the current study. Another extension would be multimodal fusion, such that images, video streams, and metadata could be ingested along with text content and support richer situational awareness. In the same vein, the incorporation of geolocation inference could permit the system to deliver location-sensitive insights and thus support more targeted intervention by emergency responders. Finally, the creation of an interactive dashboard for visualizations within a real-time manner would support decision-makers within governments and their humanitarian partners by allowing them to monitor the evolution of disasters and sentiment on the part of the public throughout the course of the events.

It not only manifests the possibility of integrating the Bidirectional LSTM and MLP models with the social media analysis for the purpose of disaster informatics but also provides an early step for the realization of worldwide deployable, multimodal, and online systems for assisting the proper disaster response and the resilience planning process.

Code and Data Availability: The datasets used in this paper are collected from three publicly available datasets: CrisisLex T6, CrisisLex T26, and CrisisNLP. The integrated dataset, along with the full source code for preprocessing, training, and testing the models, can be accessed from the following repository: <https://github.com/Saimul-Shanto/disaster-type-stage-classification>

References

1. Adams, N., Artigiani, E.E., Wish, E.D.: Choosing your platform for social media drug research and improving your keyword filter list. *Journal of drug issues* **49**(3), 477–492 (2019)
2. Alam, F., Qazi, U., Imran, M., Ofli, F.: Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. In: *Proceedings of the International AAAI Conference on Web and social media*. vol. 15, pp. 933–942 (2021)
3. Alam, F., Sajjad, H., Imran, M., Ofli, F.: Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In: *Proceedings of the International AAAI conference on web and social media*. vol. 15, pp. 923–932 (2021)
4. Albahri, A.S., Khaleel, Y.L., Habeeb, M.A., Ismael, R.D., Hameed, Q.A., Deveci, M., Homod, R.Z., Albahri, O.S., Alamoodi, A.H., Alzubaidi, L.: A systematic review of trustworthy artificial intelligence applications in natural disasters. *Computers and Electrical Engineering* **118**, 109409 (2024)
5. Dharrao, D., Aadithyanarayanan, M., Mital, R., Vengali, A., Pangavhane, M., Rajput, S., Bongale, A.M.: An efficient method for disaster tweets classification using gradient-based optimized convolutional neural networks with bert embeddings. *MethodsX* **13**, 102843 (2024)
6. Garcia-Arteaga, J., Zambrano-Zambrano, J., Parraga-Alava, J., Rodas-Silva, J.: An effective approach for identifying keywords as high-quality filters to get emergency-implicated twitter spanish data. *Computer Speech & Language* **84**, 101579 (2024)
7. Giachanou, A., Crestani, F.: Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)* **49**(2), 1–41 (2016)
8. Gurney, K.: *An introduction to neural networks*. CRC press (2018)
9. Hallegatte, S.: A cost effective solution to reduce disaster losses in developing countries: hydro-meteorological services, early warning, and evacuation. *World Bank policy research working paper* (6058) (2012)
10. Hayat, M.K., Daud, A., Alshdadi, A.A., Banjar, A., Abbasi, R.A., Bao, Y., Dawood, H.: Towards deep learning prospects: insights for social media analytics. *IEEE access* **7**, 36958–36979 (2019)
11. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* **21**(9), 1263–1284 (2009)
12. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. *ACM computing surveys (CSUR)* **47**(4), 1–38 (2015)
13. Imran, M., Mitra, P., Castillo, C.: Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894* (2016)
14. Kabir, M.Y., Madria, S.: 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*. pp. 269–278 (2019)
15. Mohammad, S., Al Jobair, A., Shaiara, F., Sharmin, S.: Bridging the gap: A case study of utilizing social media to accelerate recovery and structuring disaster management. In: *International Conference on Human-Computer Interaction*. pp. 154–172. Springer (2024)
16. Mukhtiar, W., Rizwan, W., Habib, A., Afridi, Y.S., Hasan, L., Ahmad, K.: Relevance classification of flood-related twitter posts via multiple transformers. *arXiv preprint arXiv:2301.00320* (2023)

17. Olteanu, A., Castillo, C., Diaz, F., Vieweg, S.: Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In: Proceedings of the international AAAI conference on web and social media. vol. 8, pp. 376–385 (2014)
18. Olteanu, A., Vieweg, S., Castillo, C.: 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. pp. 994–1009 (2015)
19. Phengsuwan, J., Shah, T., Thekkummal, N., Wen, Z., Sun, R., Pullarkatt, D., Thirugnanam, H., Ramesh, M., Morgan, G., James, P., et al.: Use of social media data in disaster management: A survey. *future internet*, 13 (2), 46 (2021)
20. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
21. Rodríguez-Ibáñez, M., Casáñez-Ventura, A., Castejón-Mateos, F., Cuenca-Jiménez, P.M.: A review on sentiment analysis from social media platforms. *Expert Systems with Applications* **223**, 119862 (2023)
22. Saleem, S., Mehrotra, M.: Context-aware transfer learning approach to detect informative social media content for disaster management. *International Journal of Advanced Computer Science & Applications* **15**(1) (2024)
23. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* **45**(11), 2673–2681 (1997)
24. Shanto, M.S.H., Mohammad, S., Rahman, M.S., Eti, S.A., Mia, N., Sharmin, S.: Hierarchical multimodal detection of hate in code-mixed bengali memes: An explainable human–ai approach. In: International Conference on Human-Computer Interaction. pp. 351–367. Springer (2026)
25. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: 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. pp. 1079–1088 (2010)
26. Xu, Q.A., Chang, V., Jayne, C.: A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal* **3**, 100073 (2022)