Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084547

ARTICLE

Attention-Enhanced Hybrid Deep Learning for Disaster-Related Tweet Classification

Sheraz Ali Hassan[1], Hamid Masood Khan[1], Muhammad Javed[1], Mohd Faizal Bin Yusof [2], Jamil Abedalrahim Jamil Alsayaydeh[3,*], Fida Muhammad Khan[4] and Inam Ullah[5,*]

[1]Gomal Research Institute of Computing (GRIC), Gomal University, Dera Ismail Khan, Pakistan
[2]General Education and Foundation Program, Faculty of Resilience, Rabdan Academy, 65 Al Inshirah Street, Abu Dhabi, United Arab Emirates
[3]Department of Engineering Technology, Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer (FTKEK), Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia
[4]Department of Computer Science, Qurtuba University of Science and Information Technology, Peshawar, Pakistan
[5]Department of Computer Engineering, Gachon University, Seongnam, Republic of Korea

*Corresponding Authors: Jamil Abedalrahim Jamil Alsayaydeh. Email: jamil@utem.edu.my; Inam Ullah. Email: inam@gachon.ac.kr

Received: 24 April 2026; Accepted: 22 July 2026; Published: 15 September 2026

ABSTRACT: The increasing frequency of natural and human-induced disasters has intensified the need for reliable methods to identify crisis-relevant information from Twitter/X streams. However, tweets are often short, noisy, informal, ambiguous, and context-dependent, making disaster-related tweet classification challenging. This study proposes a lightweight attention-enhanced hybrid deep learning framework for binary disaster-related tweet classification. The framework integrates CNN-based local feature extraction, recurrent contextual modeling, static pre-trained word embeddings, class weighting, training-only data augmentation, and learned neural attention. Two architectures, CNN–LSTM–Attention and CNN–BiGRU–Attention, are evaluated on the labeled Kaggle Disaster Tweets dataset using a leakage-aware protocol in which augmentation is applied only after dataset partitioning. Conventional machine-learning, standalone deep-learning, and Transformer-based baselines, including BERT, RoBERTa, DistilBERT, and CrisisBERT, are evaluated under the same experimental setting. Experimental results show that CNN–BiGRU–Attention achieves the highest accuracy of 94.81%, while CNN–LSTM–Attention provides slightly higher ROC-AUC and disaster-class recall. Compared with Transformer-based baselines, the proposed models achieve competitive performance with substantially fewer trainable parameters. These findings indicate that lightweight attention-enhanced hybrid models can provide an effective accuracy–efficiency trade-off for disaster-related Twitter/X monitoring, although broader validation across unseen crisis events, platforms, languages, and real-time streams remains necessary.

KEYWORDS: Disaster tweet classification; Twitter/X monitoring; crisis communication; hybrid deep learning; CNN–LSTM; CNN–BiGRU; learned attention; transformer baselines; data augmentation

1 Introduction

Natural and human-induced disasters generate urgent information needs for emergency responders, humanitarian organizations, and decision-makers. These events provide real-time warnings, eyewitness reports, damage reports, calls for help, and situation updates from Twitter/X [1,2]. Such user-generated content may help provide crisis awareness or assist in coordinating responses. However, tweets during crises are likely to be short, noisy, informal, ambiguous, multilingual, and context-dependent. They are

often filled with abbreviations, slang, misspellings, hashtags, figurative expressions, incomplete contexts, and vocabulary related to specific events [3,4]. These features make the automated classification of disaster-related tweets a difficult short-text classification problem. Traditional machine-learning techniques, such as Naïve Bayes, Support Vector Machines (SVM), and lexicon-based classifiers, have been widely used to analyse text in Twitter/X and crisis-related data [5,6]. While such approaches are efficient, they rely on manually designed features and simple textual representations, and they cannot model context, informality, and semantic ambiguity in crisis communication situations [7,8]. Some of these limitations are overcome by deep learning approaches that learn hierarchical representations directly from the text [9,10]. While recurrent models, including Long Short-Term Memory (LSTM) networks and Gated Recurrent Unit (GRU) networks, can model sequential relationships between tweet tokens [11–13], Convolutional Neural Networks (CNNs) are effective at identifying informative n-grams, crisis-related terms, and short semantic patterns. Hence, hybrid CNN–LSTM and CNN–GRU models are useful for the classification of tweets in disaster contexts, as relevance in such cases depends on both local lexical cues and surrounding context [14,15].

Another factor in crisis-related short-text classification is the use of text representation. One-hot encoding and bag-of-words features are examples of sparse representations that are not able to model semantic relationships between words. Dense embedding methods, such as Word2Vec and contextual embedding methods, provide more informative representations for Twitter/X text [16–19]. However, not every token has the same significance for the final classification decision. Some words, such as fire, flood, earthquake, explosion, and help, can be disaster-relevant in some contexts but may also be used metaphorically or in non-disaster contexts. Learned attention mechanisms can be used to assign adaptive weights to informative token-level representations. In this study, attention is employed as a generic learned weighting mechanism rather than an explicitly domain-informed module based on crisis lexicons, event metadata, geographic signals, temporal priors, or external knowledge sources. Recent Transformer-based and crisis-domain language models, such as BERT, RoBERTa, DistilBERT, CrisisBERT, BERTweet, and BERT-based disaster tweet classifiers, have shown excellent performance in natural language processing and crisis classification tasks [19–25]. These models provide more powerful contextual representations; however, they tend to have large parameter sizes, high memory requirements, long fine-tuning times, and greater implementation complexity than lightweight neural architectures. As a result, CNN–LSTM–Attention and CNN–BiGRU–Attention remain valuable for cases where computational efficiency, simplicity of implementation, and deployment feasibility are important. All proposed models and Transformer-based baselines are evaluated on the same labeled Kaggle Disaster Tweets dataset, train–validation–test split, preprocessing protocol, random seed setting, and evaluation procedure.

Although there have been significant advances in the classification of disaster-related tweets, some challenges remain. Many studies are based on a single benchmark dataset, which means they provide limited evidence of cross-event, cross-platform, multilingual, and temporal generalization. In some cases, data augmentation is performed before data splitting, potentially leading to train–test leakage if original tweets and their augmented counterparts are split into different datasets. Class imbalance is often recognized but is not always assessed using class-wise or threshold-independent measures. Moreover, several studies report performance gains but lack ablation analysis, repeated-run robustness assessment, confidence-interval reporting, learning-curve analysis, or parameter-efficiency comparison. It is also important to note that the learned attention described in such studies is not necessarily a domain-informed attention mechanism.

In this regard, the current work introduces a lightweight attention-enhanced hybrid deep learning system for the binary classification of disaster-related tweets. To model both local textual patterns and sequential contextual dependencies, two models are investigated: CNN–LSTM–Attention and CNN–BiGRU–Attention. The framework includes static pre-trained embeddings, data augmentation during

training, class weighting, learned neural attention, ablation analysis, repeated-run robustness evaluation, confidence-interval reporting of performance, attention-heatmap visualization, parameter-complexity analysis, and direct comparison with Transformer-based baselines. The Kaggle Disaster Tweets dataset is split before augmentation, and augmentation is performed only on the training set to mitigate leakage issues. Given that the evaluation is conducted on one benchmark dataset, the results should be taken as evidence for the dataset used and not as full evidence of robustness in cross-dataset, cross-event, multilingual, or real-time deployment contexts.

The main contributions of this study are summarized as follows:

- We develop two lightweight attention-enhanced hybrid architectures, CNN–LSTM–Attention and CNN–BiGRU–Attention, for binary disaster-related tweet classification by combining CNN-based local feature extraction with LSTM- and BiGRU-based contextual modeling.
- We introduce a leakage-aware experimental protocol in which the labeled dataset is partitioned before augmentation, and augmented samples are generated only from the training subset.
- We evaluate learned neural attention as an adaptive token-weighting component and clearly distinguish it from explicitly domain-informed attention based on crisis lexicons, metadata, geographic signals, temporal priors, or external knowledge sources.
- We directly compare the proposed models with BERT, RoBERTa, DistilBERT, and CrisisBERT under the same experimental protocol and report parameter-efficiency analysis to highlight the trainable-parameter compactness of the proposed lightweight models.

The remainder of this paper is organized as follows. Section 2 reviews related work, and Section 3 presents the proposed methodology. Section 4 reports the experimental results, while Section 5 discusses the main findings, limitations, and practical implications. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Work

Tweet classification for disaster events has emerged as one of the key research areas, as Twitter/X can provide instant awareness information during disasters. In emergencies, users can broadcast warnings, eyewitness accounts, damage reports, requests for assistance, and real-time information in brief posts. However, this type of content is sometimes verbose, unclear, colloquial, and situational, with shortened terms, slang, misspellings, hashtags, metaphors, and jargon. All these properties make disaster-related tweet classification a difficult short-text classification task. Crisis-related Twitter/X data have been annotated, filtered, and classified using conventional machine-learning approaches, as well as topic-model-based labeling, lexicon-based filtering, and human-annotated crisis corpora [26–28]. These methods work for crisis response workflows, but they rely on handcrafted features or simple textual representations. Thus, they fail to capture semantic ambiguity, contextual variation, and sequential dependency in disaster-related tweets. Some of these drawbacks can be overcome by deep learning methods that directly learn hierarchical feature representations from text. CNN-based models work well for capturing local textual patterns, such as informative n-grams, crisis-related terms, and short semantic phrases [29–31]. However, CNN-based models may not be able to capture sequential relationships across the whole tweet. Recurrent models such as LSTM, BiLSTM, and GRU can be used for noisy short Twitter/X texts to model contextual dependency between tokens [32,33]. Thus, hybrid CNN–LSTM and CNN–BiGRU architectures remain applicable to the classification task in the context of disaster-related tweets, as disaster relevance may sometimes rely on both local lexical cues and the surrounding context.

In neural classifiers, attention mechanisms have also been adopted to give more weight to informative textual or multimodal representations [34,35]. In disaster-related tweet classification, attention can highlight token-level hidden representations that contribute more strongly to the final decision. Many attention-based models use standard learned attention, where attention weights are computed from hidden-state representations. These models do not necessarily incorporate external crisis-specific knowledge, such as crisis lexicons, event metadata, geographic signals, temporal context, or keyword priors. Therefore, standard learned attention should be distinguished from explicitly domain-informed or crisis-aware attention mechanisms. Transformer-based models, including BERT, RoBERTa, DistilBERT, BERTweet, and crisis-domain Transformer variants, have achieved strong performance in natural language processing because of their contextual representation and self-attention capabilities [20,22–25]. These models can capture long-range dependencies more effectively than many earlier CNN-, LSTM-, and GRU-based architectures. BERT-based disaster tweet classifiers and contextual-embedding methods have also been explored for disaster/non-disaster tweet classification [19,21,36,37]. Nevertheless, Transformer models usually require larger parameter budgets, higher memory usage, longer training time, and careful fine-tuning. Therefore, this study evaluates BERT, RoBERTa, DistilBERT, and CrisisBERT under the same experimental protocol used for the proposed models. This comparison helps examine whether lightweight CNN–LSTM–Attention and CNN–BiGRU–Attention models can provide a favorable accuracy–efficiency trade-off.

Public crisis datasets and benchmarks have played an important role in evaluating disaster-related Twitter/X models. CrisisLex supports crisis-related microblog collection and filtering, while human-annotated Twitter/X corpora provide labeled resources for crisis-related natural language processing [27,38]. HumAID provides human-annotated disaster incident data with deep learning benchmarks, CrisisBench consolidates multiple crisis-related datasets for humanitarian information processing, and CrisisMMD extends crisis analysis to multimodal Twitter/X data from natural disasters [39–41]. These resources show that strong performance on a single dataset does not necessarily imply robust generalization across unseen crisis events, geographic contexts, languages, humanitarian categories, platforms, or real-time streams. Another key factor in disaster-related tweet classification is reproducibility. Some papers provide limited information about the train–validation–test split, tokenization process, vocabulary size, sequence length, embedding configurations, hyperparameter choices, random seeds, early stopping, class weighting, computing environment, and augmentation procedures. Data augmentation is especially sensitive, as it can lead to train–test leakage if performed before data partitioning. In such instances, one original tweet may be placed in one subset, while its transformed version may be placed in another. A better protocol is to split the original labelled dataset into the training set and the test set, and then augment only the training set. Although augmentation techniques such as synonym substitution, random deletion, random insertion, back-translation, and random word swapping can help increase linguistic diversity, they must be applied carefully because they may change the urgency, event meaning, or disaster relevance of the text.

Model assessment is also complicated by class imbalance and incomplete evaluation. Accuracy alone can be misleading if one class is more common or more operationally important than the other. In crisis monitoring, false negatives can result in crucial crisis-related information being overlooked, while false positives can lead to redundant alerts or distract from and detract from the response process. Hence, class-wise precision, recall, F1-score, ROC-AUC, confusion matrix analysis, and error interpretation, in addition to accuracy, should be reported [42]. In addition, systematic ablation and robustness analyses to assess the individual contributions of CNN feature extraction, LSTM/BiGRU contextual modeling, attention, embedding type, augmentation, and class weighting are often not conducted. Likewise, architecture comparisons that result in small performance differences should be viewed with some skepticism unless they are supported by multiple runs, confidence intervals, or statistical significance testing [43].

Previous studies have demonstrated the usefulness of deep learning and Transformer-based models for disaster-related tweet classification. However, several open issues remain, including incomplete reproducibility reporting, limited use of leakage-aware augmentation protocols, insufficient component-level ablation, limited robustness analysis, weak interpretability discussion, and inadequate consideration of deployment and ethical risks. In this paper, two lightweight attention-enhanced hybrid models, CNN–LSTM–Attention and CNN–BiGRU–Attention, are investigated for binary disaster-related tweet classification. The study focuses on the controlled and reproducible integration of CNN-based local feature extraction, LSTM/BiGRU-based contextual modeling, learned attention-based token weighting, static pre-trained embeddings, training-only augmentation, class imbalance handling, Transformer-based baseline comparison, ablation analysis, repeated-run robustness assessment, attention-assisted interpretability analysis, confidence-interval reporting, and multi-metric evaluation. The proposed models are therefore positioned as leakage-aware, reproducible, and resource-conscious applied classification approaches for disaster-related Twitter/X monitoring.

3 Proposed Methodology

This study proposes two lightweight attention-enhanced hybrid architectures, CNN–LSTM–Attention and CNN–BiGRU–Attention, for binary disaster-related tweet classification. The framework integrates CNN-based local feature extraction, LSTM/BiGRU-based contextual modeling, static pre-trained word embeddings, learned neural attention, class imbalance handling, and training-only augmentation. The objective is to classify each tweet as disaster-related or non-disaster-related under a leakage-aware and reproducible evaluation protocol. Conventional machine-learning, standalone deep-learning, and Transformer-based baselines are evaluated using the same dataset split and evaluation protocol. The Transformer-based baselines include BERT, RoBERTa, DistilBERT, and CrisisBERT. The methodological contribution is the controlled integration of CNN-based local pattern extraction, recurrent contextual learning, learned attention-based token weighting, and training-only augmentation. CNN–LSTM–Attention and CNN–BiGRU–Attention are based on well-known hybrid classification approaches; therefore, the proposed approach is not a new neural architecture, but rather a reproducible applied classification method. This attention mechanism is standard learned neural attention, without explicitly incorporating any crisis lexicons, keyword priors, event metadata, geographic signals, temporal context, or external domain knowledge. Therefore, the proposed models are described as attention-enhanced rather than explicitly crisis-aware.

Fig. 1 illustrates the structure of the proposed CNN–LSTM–Attention and CNN–BiGRU–Attention framework for disaster-related tweet classification.

3.1 Dataset, Preprocessing, and Augmentation

This study uses the Kaggle Disaster Tweets dataset for supervised disaster-related tweet classification. The complete Kaggle competition data contains 10,876 tweet instances, including 7613 labeled training tweets and 3263 unlabeled official test tweets. Since ground-truth labels are not available for the official test set, only the 7613 labeled tweets are used for supervised training, validation, and testing. The labeled subset consists of 3271 disaster-related tweets and 4342 non-disaster tweets. To minimize train-test leakage, the original labeled dataset is partitioned before augmentation. Stratified sampling is used to divide the labeled data into an 80% training-development set and a 20% independent test set. From the training-development subset, 10% is reserved for validation. Augmentation is then applied only to the final training set, while the validation and test sets remain unchanged. This protocol prevents augmented variants of the same original tweet from appearing across the training, validation, and testing partitions.

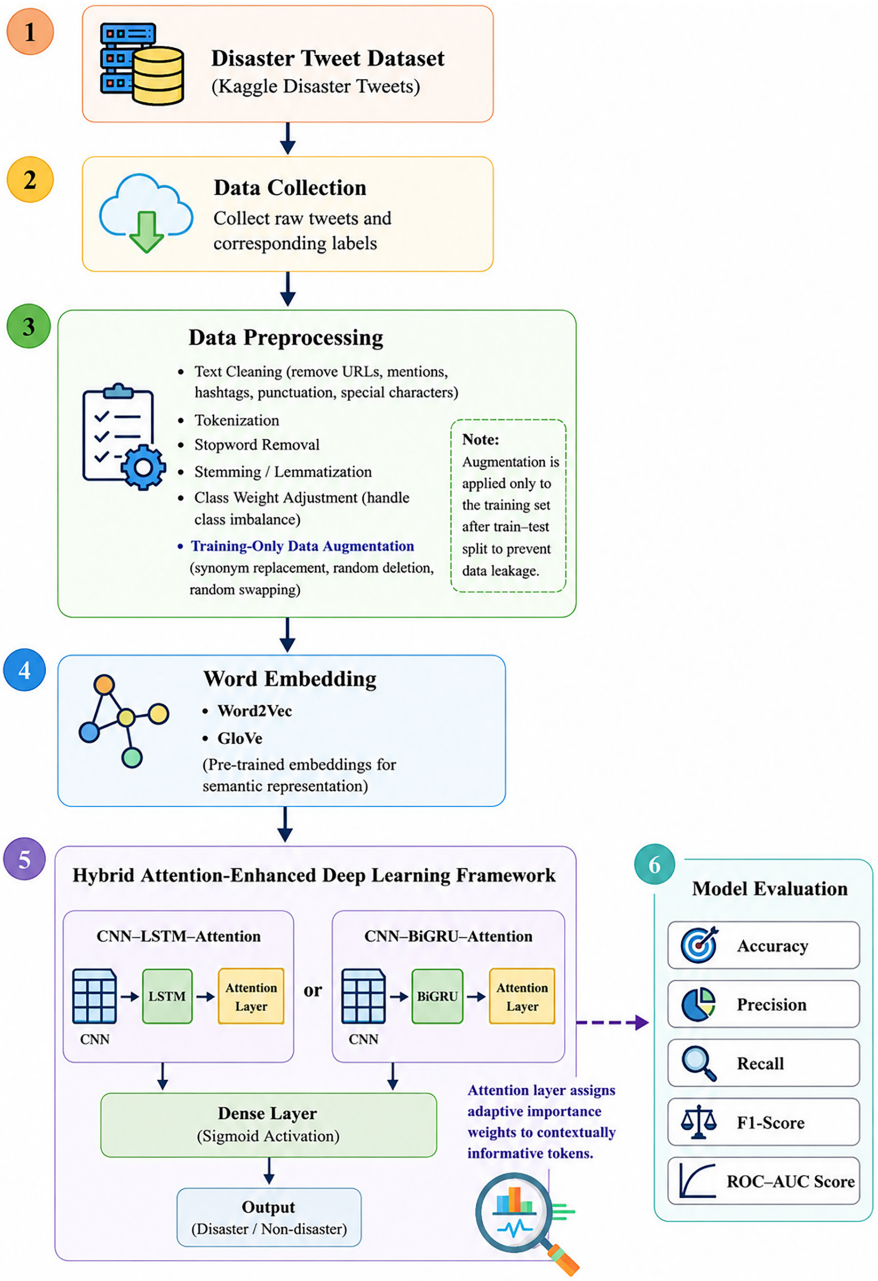

Figure 1: Architecture of the proposed CNN–LSTM–attention and CNN–BiGRU–attention framework, comprising static word embeddings, CNN-based local feature extraction, recurrent contextual modeling, learned token-level attention, and sigmoid-based binary classification.

Table 1 reports the dataset distribution under the controlled training-only augmentation protocol.

The preprocessing pipeline includes URL removal, user-mention removal, punctuation and symbol cleaning, lower casing, tokenization, word normalization, sequence padding, and embedding mapping. Stopword removal is not performed extensively to reduce non-discriminative noise while retaining disaster-relevant semantic clues. Since the dataset has moderate class imbalance, class weights are used during training to reduce majority-class bias. Training-only augmentation is performed using synonym replacement, random deletion, and random word swapping. The original training samples are retained together with

their augmented variants, increasing the effective training size by a factor of four. Validation and testing samples are not augmented. Although this strategy improves training diversity while reducing leakage risk, augmentation-specific ablation is not conducted in this study and is reserved for future work.

Table 1: Dataset distribution under controlled training-only augmentation.

Category	Original Samples	Training Samples	Validation Samples	Training after Augmentation	Testing Samples
Disaster tweets	3271	2355	262	9420	654
Non-disaster tweets	4342	3126	347	12,504	869
Total	7613	5481	609	21,924	1523

3.2 Embedding Representation

Each tweet is represented as a sequence of dense word embeddings. Given a tweet sequence $S = w_1, w_2, \ldots, w_T$, each token w_i is mapped into an embedding vector $e_i \in \mathbb{R}^d$:

$$e_i = E(w_i), \tag{1}$$

where $E(\cdot)$ denotes the pre-trained embedding lookup function and d is the embedding dimension. The complete tweet representation is defined as:

$$X = [e_1, e_2, \ldots, e_T] \in \mathbb{R}^{T \times d}, \tag{2}$$

where T denotes the maximum sequence length. Word2Vec 300-dimensional and GloVe 100-dimensional embeddings are considered for the proposed lightweight hybrid models. The reported architecture summaries use the GloVe 100-dimensional configuration with a 30,000-token vocabulary, resulting in 3,000,000 non-trainable embedding parameters.

For Transformer-based baselines, model-specific subword tokenization is used instead of static word-embedding lookup. BERT, RoBERTa, DistilBERT, and CrisisBERT generate contextual representations through their pre-trained Transformer encoders and are fine-tuned under the same train–validation–test split and evaluation protocol.

3.3 Proposed CNN–LSTM–Attention and CNN–BiGRU–Attention Models

The proposed framework contains two hybrid architectures: CNN–LSTM–Attention and CNN–BiGRU–Attention. Both models first apply a CNN layer to extract local textual features from embedded tweet sequences. The convolutional feature map is computed as:

$$F_{\text{cnn}} = \text{Conv1D}(X, f, k), \tag{3}$$

where F_{cnn} denotes the CNN feature map, X is the embedded tweet sequence, f is the number of filters, and k is the kernel size.

In CNN–LSTM–Attention, the CNN feature map is passed to an LSTM layer to capture sequential contextual dependencies:

$$H_{\text{lstm}} = \text{LSTM}(F_{\text{cnn}}, u), \tag{4}$$

where H_{lstm} denotes the LSTM hidden representation and u represents the number of LSTM units.

In CNN–BiGRU–Attention, the CNN feature map is passed to a bidirectional GRU layer to capture contextual information from both forward and backward directions:

$$H_{\text{bigru}} = \text{BiGRU}\left(F_{\text{cnn}}, u\right), \tag{5}$$

where H_{bigru} denotes the bidirectional contextual representation generated by the BiGRU layer. Compared with CNN–LSTM–Attention, CNN–BiGRU–Attention captures both preceding and subsequent contextual cues, although it introduces higher trainable-parameter complexity.

3.4 Learned Attention Mechanism

A learned attention mechanism is applied after the LSTM or BiGRU layer to assign adaptive weights to informative token-level hidden states. Let h_i denote the hidden representation of the i^{th} token. The unnormalized attention score is computed as:

$$s_i = v^T \tanh\left(W_h h_i + b_h\right), \tag{6}$$

where W_h, b_h, and v are trainable attention parameters. The normalized attention weight is obtained using the softmax function:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{T} \exp(s_j)}. \tag{7}$$

The final attention-based context vector is computed as:

$$H_{\text{att}} = \sum_{i=1}^{T} \alpha_i h_i. \tag{8}$$

The context vector H_{att} is passed to a dense sigmoid output layer for final binary classification. Because the attention mechanism is learned solely from internal hidden representations and does not explicitly incorporate crisis dictionaries or lexicons, keyword priors, event metadata, geographic signals, temporal priors, or external domain knowledge, the proposed models are described as attention-enhanced rather than explicitly crisis-aware. The learned attention weights serve only as auxiliary interpretive signals indicating the relative importance of token-level hidden representations and should not be interpreted as causal attributions or as a complete explanation of the models' behavior.

3.5 Training Configuration

Both proposed models are trained using the same data split, preprocessing pipeline, embedding configuration, class weighting, early stopping strategy, and evaluation metrics. Conventional machine-learning baselines, including SVM and Naive Bayes, are evaluated using TF-IDF features. Standalone deep-learning baselines, including CNN and LSTM, are evaluated using tokenized, padded, embedding-based input sequences. Transformer-based baselines, including BERT, RoBERTa, DistilBERT, and CrisisBERT, are evaluated using their corresponding model-specific subword tokenizers and fine-tuning procedures under the same train–validation–test split, random seed setting, and evaluation protocol.

Table 2 summarizes the main experimental configuration.

Table 2: Experimental configuration.

Component	Setting
Dataset	Kaggle Disaster Tweets labeled subset
Task	Binary classification: disaster vs. non-disaster tweet
Training/validation/testing	5481/609/1523 samples
Training after augmentation	21,924 samples
Split protocol	Stratified 80:20 split; 10% validation from training-development subset
Augmentation	Training-only synonym replacement, random deletion, and random word swapping
Classical ML input	TF-IDF features
Neural model input	Tokenized and padded embedding-based sequences
Transformer model input	Model-specific subword tokenization
Tokenizer	Keras Tokenizer for CNN/LSTM-based models; model-specific tokenizer for Transformer baselines
Maximum vocabulary size	30,000 tokens
Maximum sequence length	40 tokens
Embeddings	Word2Vec 300-dimensional and GloVe 100-dimensional for proposed lightweight models
Transformer baselines	BERT, RoBERTa, DistilBERT, and CrisisBERT
Optimizer	Adam
Learning rate	0.001
Loss function	Binary cross-entropy
Batch size	64
CNN filters/kernel size	128/3
Recurrent units	128
Dropout rate	0.3
Early stopping	Validation loss; patience = 3 epochs
Random seed	42
Framework	TensorFlow 2.12 and Keras 2.12
Hardware	NVIDIA RTX 3090 GPU, Intel i9 CPU, 64 GB RAM

A fixed random seed of 42 is used to support reproducible dataset partitioning and model training. Repeated-run robustness is also reported to assess the stability of the proposed models across independent executions.

Binary cross-entropy is used as the training objective:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\widehat{y}_i) + (1 - y_i)\log(1 - \widehat{y}_i)\right], \tag{9}$$

where N is the number of training samples, y_i is the ground-truth label, and $\widehat{y}_i$ is the predicted probability. Adam optimization is used for model training, while dropout regularization and early stopping help reduce overfitting. Class weighting is applied to address class imbalance by assigning greater importance to minority-class errors and reducing majority-class bias.

3.6 Training Procedure and Evaluation Protocol

Algorithm 1 summarizes the shared training and evaluation procedure used for CNN–LSTM–Attention and CNN–BiGRU–Attention. Transformer-based baselines are fine-tuned separately using their own subword tokenizers and pre-trained encoder architectures, but they are evaluated using the same dataset split and evaluation metrics.

Algorithm 1: Training procedure for CNN–LSTM–attention and CNN–BiGRU–attention

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N},\ y_i \in \{0,1\}$
$(\mathcal{D}_{tr}, \mathcal{D}_{val}, \mathcal{D}_{te}) \leftarrow \text{StratifiedSplit}(\mathcal{D})$
$\widetilde{\mathcal{D}}_{tr} \leftarrow \mathcal{D}_{tr} \cup \mathcal{A}(\mathcal{D}_{tr})$
$\mathcal{A} = \{\text{SynRep}, \text{RandDel}, \text{RandSwap}\}$
for $(x_i, y_i) \in \widetilde{\mathcal{D}}_{tr}$ **do**
    $S_i = \text{Pad}(\text{Tok}(\text{Clean}(x_i)), T)$
    $X_i = [E(w_1), E(w_2), \ldots, E(w_T)] \in \mathbb{R}^{T \times d}$
end for
$w_c = \dfrac{N}{2N_c}, \qquad c \in \{0,1\}$
for $m \in \{\text{CNN-LSTM-Att}, \text{CNN-BiGRU-Att}\}$ **do**
    $F_i = \text{MaxPool}\left(\sigma(\text{Conv1D}(X_i; f, k))\right)$
    if m = CNN-LSTM-Att **then**
        $H_i = \text{LSTM}(F_i; u)$
    else
        $H_i = \text{BiGRU}(F_i; u)$
    end if
    $s_{it} = v^T \tanh(W_h h_{it} + b_h)$
    $\alpha_{it} = \dfrac{\exp(s_{it})}{\sum_{j=1}^{T} \exp(s_{ij})}$
    $z_i = \sum_{t=1}^{T} \alpha_{it} h_{it}$
    $\widehat{y}_i = \sigma(W_o z_i + b_o)$
    $\mathcal{L}(\theta) = -\frac{1}{|\widetilde{\mathcal{D}}_{tr}|} \sum_{i=1}^{|\widetilde{\mathcal{D}}_{tr}|} w_{y_i} \left[y_i \log(\widehat{y}_i) + (1 - y_i) \log(1 - \widehat{y}_i)\right]$
    $\theta^{(r+1)} \leftarrow \text{Adam}\left(\theta^{(r)}, \nabla_\theta \mathcal{L}(\theta^{(r)}), \eta\right)$
    $\theta_m^* \leftarrow \arg\min_\theta \mathcal{L}_{val}(\theta)$
    $\widehat{\mathbf{y}}_{te}^{m} \leftarrow \mathbb{I}\left(\sigma(f_{\theta_m^*}(\mathcal{D}_{te})) \geq 0.5\right)$
    $\mathcal{M}_m \leftarrow \{\text{Acc}, \text{Prec}, \text{Rec}, \text{F1}, \text{AUC}, \text{CM}\}$
end for
$m^* \leftarrow \arg\max_m \left(\text{F1}_m, \text{AUC}_m, \text{Acc}_m\right)$

The models are evaluated using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix analysis, learning curves, ablation analysis, parameter-complexity analysis, and repeated-run robustness. Since disaster-related tweet classification is operationally sensitive, class-wise recall and false-negative behavior are interpreted alongside overall accuracy.

4 Experimental Results

This section evaluates the proposed CNN–LSTM–Attention and CNN–BiGRU–Attention models on the labeled Kaggle Disaster Tweets dataset under the leakage-aware protocol described in Section 3. Performance is assessed using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices, learning curves, ablation analysis, repeated-run robustness, confidence intervals, and parameter-efficiency comparison. The results reported in this section primarily reflect the effectiveness of the proposed models on this specific benchmark dataset. Their generalization ability across unseen crisis events, social-media platforms, multiple languages, and real-time streaming data requires further verification.

4.1 Dataset Distribution

Fig. 2 shows the class distribution of the labeled Kaggle Disaster Tweets dataset. The dataset contains 3271 disaster-related tweets and 4342 non-disaster tweets, indicating moderate class imbalance. Therefore, class weighting is applied during training, and class-wise metrics are reported together with overall accuracy. These results interpreted as dataset-specific evidence, and further validation is required across unseen crisis events, other platforms, multilingual Twitter/X streams, and real-time deployment scenarios.

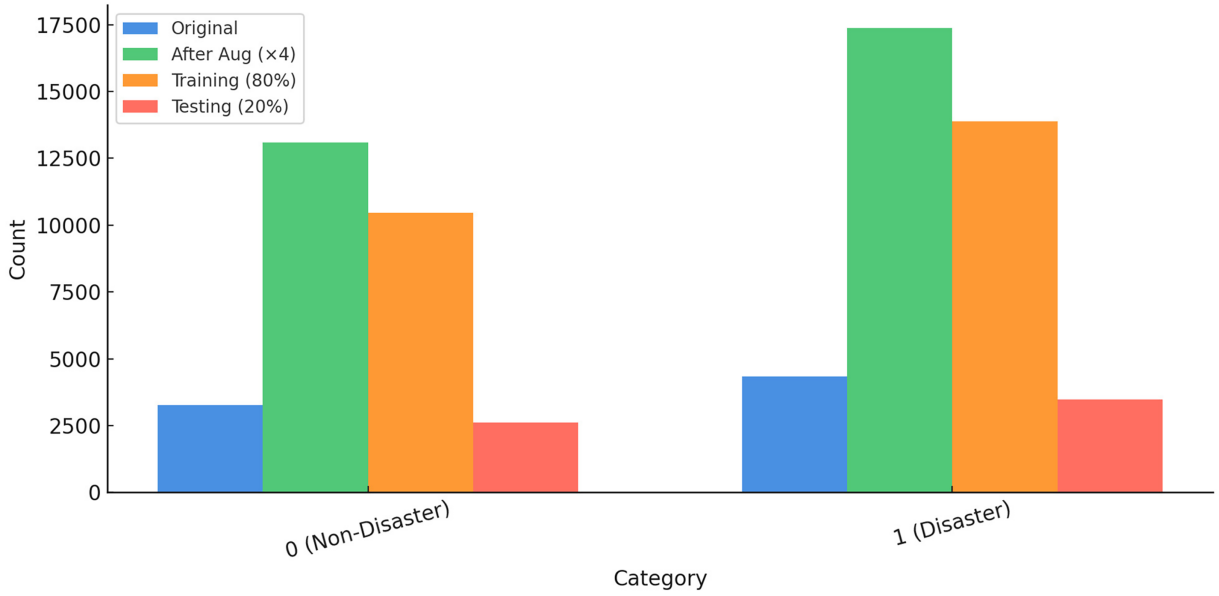

Figure 2: Class distribution of the labeled Kaggle disaster tweets dataset, showing 3271 disaster-related and 4342 non-disaster tweets and the moderate class imbalance considered during model training and evaluation.

4.2 Overall Performance Comparison

Table 3 compares the proposed models with conventional machine-learning and standalone deep-learning baselines under the same evaluation protocol. SVM and Naive Bayes are evaluated using TF-IDF features, whereas CNN, LSTM, CNN–LSTM–Attention, and CNN–BiGRU–Attention are evaluated using tokenized embedding-based sequences.

As shown in Table 3, both proposed models outperform the conventional machine-learning and standalone neural baselines. CNN–BiGRU–Attention achieves the highest accuracy of 94.81%, followed by CNN–LSTM–Attention with 94.39%. However, the absolute difference between the two proposed models is

only 0.42 percentage points; therefore, this improvement is interpreted as marginal rather than as evidence of decisive architectural superiority.

Table 3: Accuracy comparison of proposed and baseline models.

Model	Input Representation	Accuracy (%)
SVM	TF-IDF	83.40
Naive Bayes	TF-IDF	81.20
CNN	Embedding sequence	89.00
LSTM	Embedding sequence	87.50
CNN–LSTM–Attention	Embedding sequence	94.39
CNN–BiGRU–Attention	Embedding sequence	94.81

Fig. 3 visually compares the accuracy of the conventional machine-learning, standalone deep-learning, and proposed attention-enhanced hybrid models.

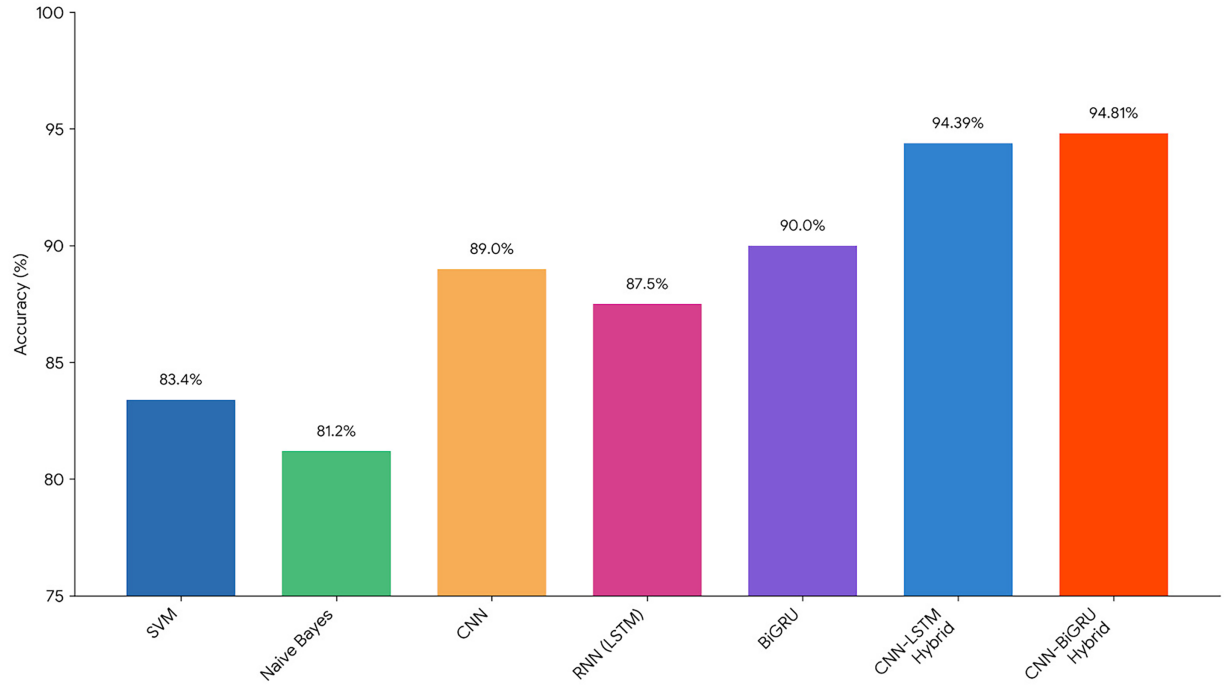

Figure 3: Accuracy comparison of the proposed CNN–LSTM–attention and CNN–BiGRU–attention models with conventional machine-learning and standalone deep-learning baselines under the same evaluation protocol.

4.3 Class-Wise and Threshold-Independent Evaluation

Table 4 reports class-wise and threshold-independent evaluation metrics for the two proposed models. CNN–BiGRU–Attention obtains higher overall accuracy, F1-score, and disaster-class precision, whereas CNN–LSTM–Attention achieves higher disaster-class recall and ROC-AUC. This indicates an accuracy–recall trade-off: CNN–BiGRU–Attention is more conservative in predicting disaster-related tweets, while CNN–LSTM–Attention is more sensitive to disaster-related instances.

Table 4: Class-wise and threshold-independent evaluation of the proposed models.

Metric	CNN–LSTM–Attention	CNN–BiGRU–Attention
Accuracy (%)	94.39	94.81
F1-score	0.930	0.940
Precision (Class 0)	0.950	0.940
Precision (Class 1)	0.930	0.960
Recall (Class 0)	0.950	0.970
Recall (Class 1)	0.940	0.920
ROC-AUC	0.986	0.983

4.4 Model-Specific Visual Analysis

Figs. 4–9 present model-specific visual evaluations for CNN–LSTM–Attention and CNN–BiGRU–Attention. The confusion matrices show class-wise prediction behavior, the ROC curves illustrate threshold-independent separability, and the learning curves report training and validation accuracy/loss trends across epochs. The confusion matrices show that CNN–BiGRU–Attention produces fewer false-positive disaster predictions than CNN–LSTM–Attention, but has a slightly higher false-negative disaster prediction rate. This is consistent with its higher precision and lower recall for the disaster class.

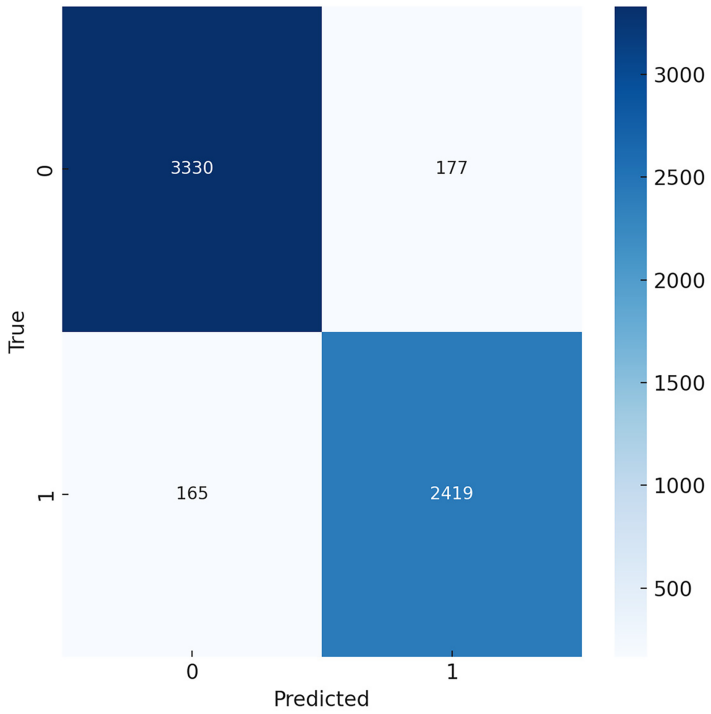

Figure 4: Confusion matrix of the CNN–LSTM–attention model on the independent test set, showing correct classifications and false-positive and false-negative predictions for disaster-related and non-disaster tweets.

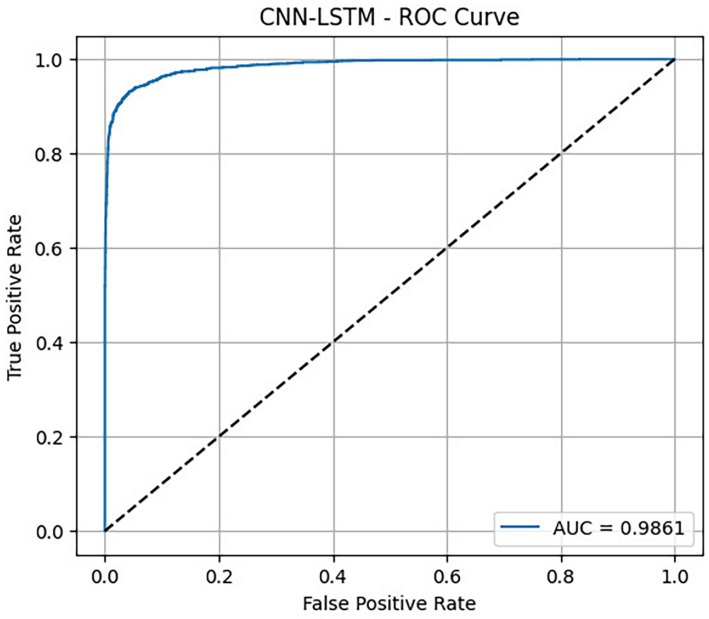

Figure 5: ROC curve of the CNN–LSTM–attention model on the independent test set, showing threshold-independent discrimination between disaster-related and non-disaster tweets with an ROC-AUC of 0.986.

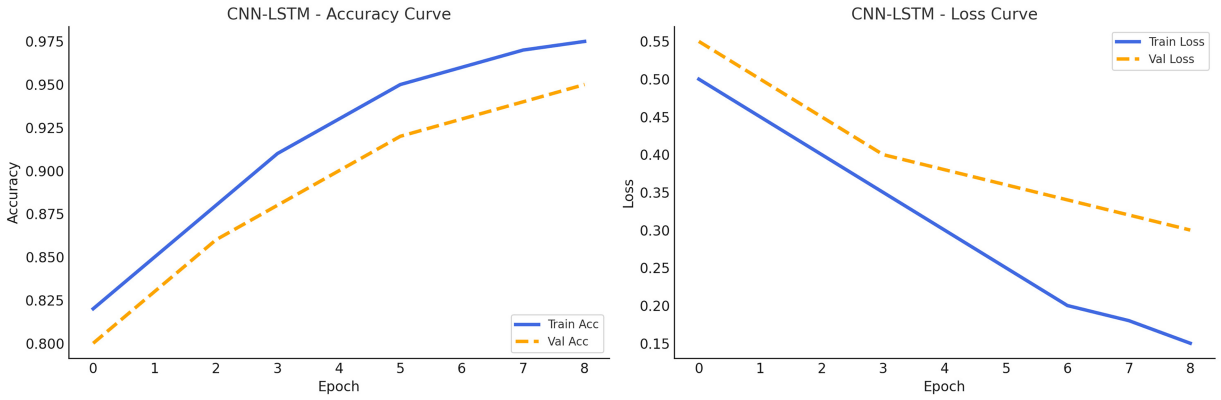

Figure 6: Training and validation accuracy and binary cross-entropy loss curves of the CNN–LSTM–attention model across epochs, illustrating model convergence and potential overfitting.

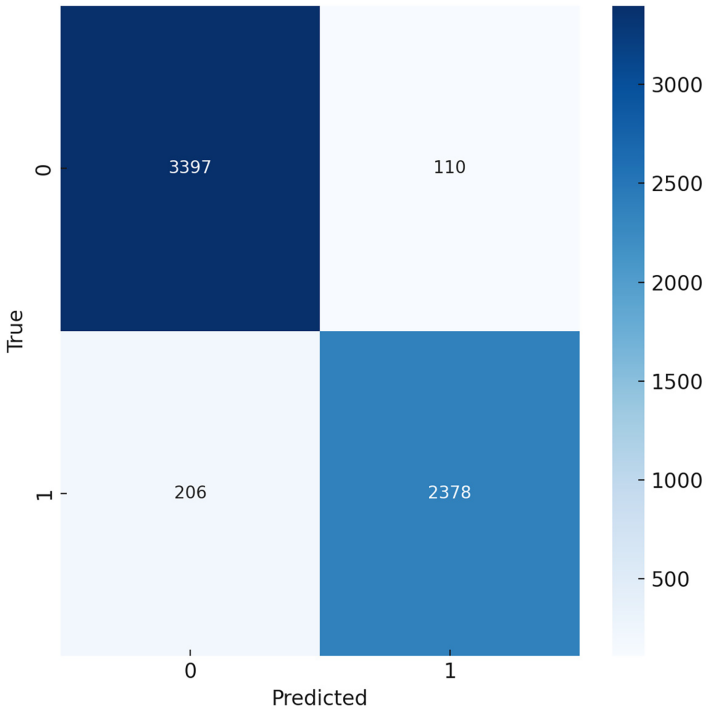

Figure 7: Confusion matrix of the CNN–BiGRU–attention model on the independent test set, showing correct classifications and false-positive and false-negative predictions for disaster-related and non-disaster tweets.

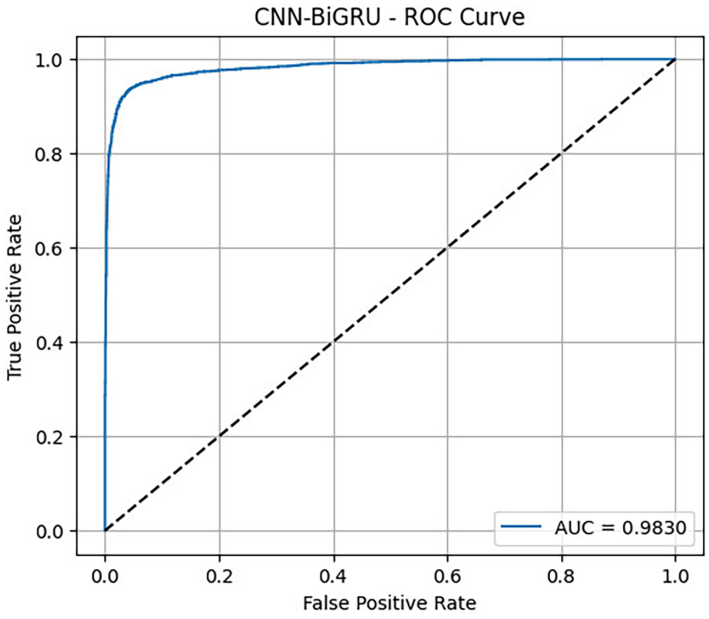

Figure 8: ROC curve of the CNN–BiGRU–attention model on the independent test set, showing threshold-independent discrimination between disaster-related and non-disaster tweets with an ROC-AUC of 0.983.

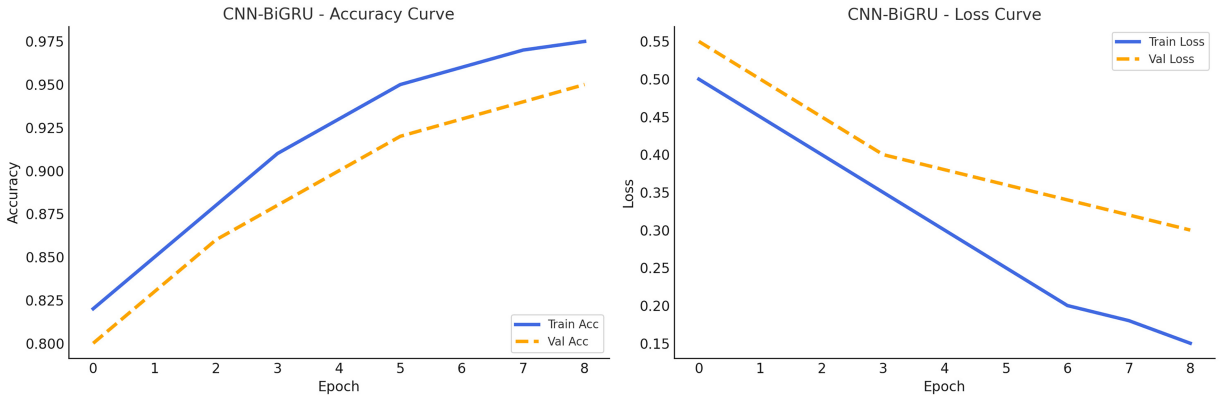

Figure 9: Training and validation accuracy and binary cross-entropy loss curves of the CNN–BiGRU–attention model across epochs, illustrating model convergence and potential overfitting.

Both proposed models show good class separability in the ROC curves. The learning curves also show stable convergence, as the validation accuracy stabilizes after a few epochs and the validation loss remains under control. This indicates that dropout, early stopping, and validation monitoring are beneficial in alleviating overfitting, despite the use of augmented training data.

4.5 Attention-Based Interpretability

Attention heatmaps for a representative set of disaster- and non-disaster-related tweets are shown in Fig. 10. In the disaster case, higher attention weights are assigned to tokens that indicate disaster relevance, such as 'fire', 'help', and 'evacuating'. In the non-disaster case, higher attention weights are assigned to tokens that support a non-disaster interpretation, such as 'song', 'love', and 'tonight'. The attention heatmaps provide auxiliary interpretive information and indicate the relative importance assigned to token-level hidden representations. These weights should not be interpreted as causal attributions or regarded as an exhaustive explanation of the model's behavior. In addition, the learnable attention mechanism does not explicitly use crisis dictionaries or lexicons, geographic cues, temporal priors, event metadata, keyword priors, or external crisis-domain knowledge. Thus, the heatmaps provide only supplementary interpretive evidence, and the proposed mechanism remains attention-enhanced rather than explicitly crisis-aware.

This heatmap analysis shows that the learned attention layer does not focus only on individual disaster-related words. Rather, it assigns higher weights to token-level hidden representations based on their contextual contribution. The attention weights, however, are not causal explanations of model behavior. They serve as an auxiliary means of interpretation and do not provide a complete explanation mechanism. Furthermore, the attention mechanism is learned from hidden representations and is not based on any explicit crisis lexicon, event metadata, temporal signals, geographic cues, or external knowledge. Thus, the heatmaps are regarded as supportive evidence of interpretability, not as complete causal explanations.

4.6 Comparison with Transformer-Based Baselines

The proposed models are compared with current NLP models, including BERT, RoBERTa, DistilBERT, and CrisisBERT, which are tested using the same train–validation–test split and evaluation protocol. Accuracy, total number of parameters, trainable parameters, and the reduction in parameters compared with BERT are given in Table 5. This comparison evaluates whether the lightweight hybrid models can provide a competitive accuracy-efficiency trade-off compared with larger Transformer-based models.

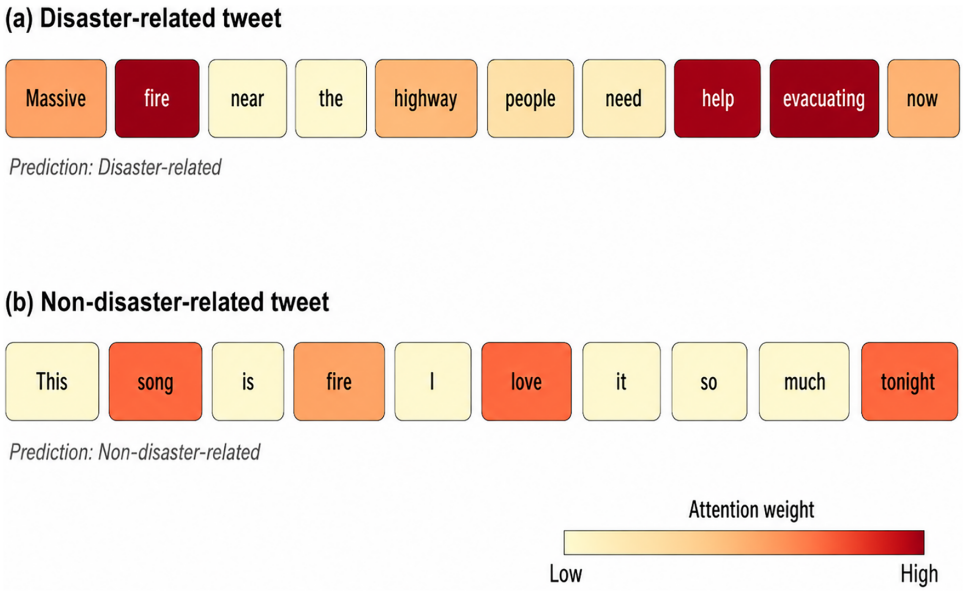

Figure 10: Attention heatmaps for representative tweets: (**a**) a disaster-related tweet and (**b**) a non-disaster tweet. The heatmaps show the relative weights assigned to token-level hidden representations.

Table 5: Performance and parameter-efficiency comparison with Transformer-based baselines.

Model	Type	Acc. (%)	Total Params	Trainable Params	Reduction vs. BERT	Deployment Implication
BERT	Transformer	93.13	~110M	~110M	Reference	High fine-tuning and memory cost
RoBERTa	Transformer	92.00	~125M	~125M	–	Higher parameter cost than BERT
DistilBERT	Lightweight Transformer	94.11	~66M	~66M	~40.00%	Lower cost than BERT
CrisisBERT	Crisis-domain Transformer	94.25	~110M	~110M	~0.00%	Crisis-specific but BERT-level cost
CNN–LSTM–Attention	Proposed lightweight hybrid	94.39	3.19M	0.19M	~99.83%	Compact trainable architecture
CNN–BiGRU–Attention	Proposed lightweight hybrid	94.81	3.30M	0.30M	~99.72%	Best accuracy with low trainable cost

The relative trainable-parameter reduction is computed as:

$$R_p = \left(1 - \frac{P_{\text{model}}}{P_{\text{BERT}}}\right) \times 100, \tag{10}$$

where P_{model} is the number of trainable parameters of the evaluated model and P_{BERT} is the number of trainable parameters in BERT.

CNN–BiGRU–Attention achieves the highest accuracy, while CNN–LSTM–Attention obtains slightly higher ROC-AUC and disaster-class recall. The proposed models require approximately 99.83% and

99.72% fewer trainable parameters than BERT, respectively. These results indicate a favorable accuracy-efficiency trade-off for resource-limited disaster-monitoring applications. However, profiling of training time, inference latency, throughput, memory consumption, and energy use is deferred to future deployment-level evaluation.

4.7 Ablation, Robustness, and Complexity Analysis

Table 6 presents the component-level ablation study. The CNN–LSTM and CNN–BiGRU variants outperform the standalone CNN and LSTM models, indicating that local feature extraction and contextual sequence modeling provide complementary benefits. Adding the attention layer further improves performance, suggesting that adaptive token-level weighting contributes positively to classification effectiveness.

Table 6: Component-level ablation analysis of the proposed framework.

Model Configuration	Accuracy (%)
CNN only	89.00
LSTM only	87.50
CNN–LSTM without attention	92.84
CNN–BiGRU without attention	93.21
CNN–LSTM–Attention	94.39
CNN–BiGRU–Attention	94.81

Table 7 defines the scope of the ablation analysis. The experiments isolate the main architectural components of the proposed models, while embedding strategies, class-balancing alternatives, augmentation variants, hyperparameter sensitivity, and alternative attention mechanisms are kept controlled rather than exhaustively varied. These factors are therefore left for future extended ablation studies.

Table 7: Scope of ablation and controlled design factors.

Design Factor	Treatment	Interpretation
CNN feature extraction	CNN-only and CNN-based hybrid variants are evaluated	Directly ablated
Recurrent modeling	LSTM-only, CNN–LSTM, and CNN–BiGRU variants are evaluated	Directly ablated
Learned attention	Attention and non-attention variants are compared	Directly ablated
Embedding, balancing, and augmentation	Static embeddings, class weights, and training-only augmentation are fixed	Controlled but not exhaustively isolated
Hyperparameters and attention variants	One fixed configuration and standard learned attention are used	Sensitivity analysis left for future work

Table 8 reports training behavior and repeated-run robustness. Both models show stable convergence and low standard deviation across five independent runs.

Table 8: Training behavior and repeated-run robustness of the proposed models.

Model	Epochs	Final Accuracy Train/Val (%)	Final Loss Train/Val	Repeated Runs Accuracy Mean ± SD
CNN–LSTM–Attention	10	98.07/94.12	0.0496/0.1775	94.39 ± 0.31
CNN–BiGRU–Attention	10	97.86/94.27	0.0571/0.1830	94.81 ± 0.27

Table 9 reports the 95% confidence intervals computed from five independent runs using the Student's t distribution. The intervals partially overlap, indicating that the improvement of CNN–BiGRU–Attention over CNN–LSTM–Attention should be interpreted as a marginal gain rather than statistically decisive superiority.

Table 9: Repeated-run accuracy with 95% confidence intervals.

Model	Mean Accuracy (%)	SD	95% Confidence Interval
CNN–LSTM–Attention	94.39	0.31	94.01–94.77
CNN–BiGRU–Attention	94.81	0.27	94.47–95.15

Table 10 reports the parameter complexity of the proposed architectures. CNN–BiGRU–Attention uses more trainable parameters because bidirectional recurrent modeling learns forward and backward contextual representations. Therefore, model selection should consider not only accuracy but also ROC-AUC, disaster-class recall, robustness, and computational cost.

Table 10: Parameter complexity of the proposed architectures.

Model	Total Params	Trainable Params	Non-Trainable Params
CNN–LSTM–Attention	3,186,881	186,881	3,000,000
CNN–BiGRU–Attention	3,302,977	302,977	3,000,000

The results show that the proposed lightweight attention-enhanced hybrid models achieve competitive accuracy with substantially fewer trainable parameters than Transformer-based baselines. However, the evaluation remains limited to a single benchmark dataset. Broader cross-dataset validation, exhaustive sensitivity analysis, extended statistical testing, and deployment-level profiling are therefore required in future work.

5 Discussion

The experimental results show that the proposed CNN–LSTM–Attention and CNN–BiGRU–Attention models achieve competitive performance for disaster-related tweet classification. CNN–BiGRU–Attention obtains the highest accuracy of 94.81%, followed by CNN–LSTM–Attention with 94.39%. Among the Transformer-based baselines, CrisisBERT achieves 94.25%, DistilBERT 94.11%, BERT 93.13%, and RoBERTa 92.00%. These results suggest that the proposed lightweight hybrid models provide a good balance between prediction performance and parameter efficiency. The improvement of CNN–BiGRU–Attention over CNN–LSTM–Attention is comparatively small and should therefore be interpreted with caution. CNN–BiGRU–Attention achieves higher accuracy and disaster-class precision, while CNN–LSTM–Attention

achieves a slightly higher ROC-AUC and disaster-class recall. This trade-off is relevant for crisis monitoring applications, where false negatives can result in missing tweets that are relevant to an emergency. Model selection should not be based on accuracy alone but should also take into account other metrics such as recall, ROC-AUC, robustness, false-negative behavior, and deployment constraints. The leakage-aware augmentation protocol enhances the reliability of the evaluation. In this study, the original labelled data are split into three sets: training, validation, and testing. Augmentation is applied only to the training set. Under this protocol, augmented versions of the same tweet are less likely to appear in different data partitions, which reduces the risk of overestimating performance. However, the effect of each augmentation component, such as synonym replacement, random deletion, and random word swapping, is not investigated separately and remains a potential area for future ablation analysis.

The comparison with Transformer-based baselines reinforces the practical value of the proposed models. Transformer-based models are effective at capturing context, but they have significantly larger trainable-parameter budgets. CNN–LSTM–Attention and CNN–BiGRU–Attention, on the other hand, achieve competitive or better accuracy compared with the Transformer variants while using significantly fewer trainable parameters. This makes them attractive for resource-constrained disaster monitoring environments where low memory usage, reduced training complexity, and deployment simplicity are important factors.

The learned attention mechanism enables adaptive token-level weighting and offers some auxiliary interpretability. The attention heatmaps show that, in general, the models assign greater attention to token-level hidden representations that are contextually relevant. However, these attention weights should not be interpreted as causal attributions or as a complete explanation of the models' behavior. The mechanism is learned without relying on explicit crisis dictionaries or lexicons, keyword priors, crisis-event metadata, geographical signals, temporal priors, or external domain knowledge. Therefore, the proposed mechanism is more accurately described as attention-enhanced rather than explicitly crisis-aware. This assessment remains benchmark-based. The experiments are conducted only on the annotated Kaggle Disaster Tweets dataset, thereby limiting the evidence of cross-event, cross-domain, multilingual, and temporal generalization. Crisis-related social media datasets may differ in event type, language, annotation scheme, geographic context, humanitarian category, modality, and temporal distribution. Therefore, future evaluations should include additional crisis datasets, such as CrisisLex, CrisisBench, HumAID, and CrisisMMD, to assess robustness under diverse crisis-monitoring conditions.

The present study has several limitations. First, the evaluation uses a single public benchmark dataset. Second, Transformer baselines are included, but exhaustive Transformer hyperparameter tuning is not performed. Third, the ablation study focuses mainly on architectural components and does not fully isolate embedding strategies, class-balancing alternatives, augmentation variants, hyperparameter sensitivity, or alternative attention mechanisms. Fourth, training time, inference latency, throughput, GPU/CPU memory usage, energy consumption, and real-time streaming performance are not measured. Finally, the proposed models should be treated as assistive screening tools under human supervision because false positives and false negatives may affect crisis-response workflows.

6 Conclusion

This study proposed two lightweight attention-enhanced hybrid models, CNN–LSTM–Attention and CNN–BiGRU–Attention, for binary disaster-related tweet classification. The proposed framework consists of a CNN-based local feature extractor, an LSTM/BiGRU-based contextual model, static word embeddings, learned neural attention, class weighting, and training-only augmentation under a leakage-aware evaluation setting. Based on the experiments, CNN–BiGRU–Attention achieves the best accuracy of 94.81%,

followed by CNN–LSTM–Attention at 94.39%. CNN–LSTM–Attention delivers slightly higher ROC-AUC and disaster-class recall. The results show that the proposed models can achieve competitive performance with significantly fewer trainable parameters compared with Transformer-based baselines, making them suitable for resource-constrained disaster monitoring applications. Although these results are promising, they should be viewed in the context of a single dataset and a controlled experimental setting. The study does not provide full evidence of cross-dataset, cross-event, multilingual, or real-time deployment robustness. In addition, wider Transformer benchmarking, factor-wise ablation analysis, more extensive statistical testing, and deployment-level efficiency profiling are still needed. Future research will extend the evaluation beyond a single dataset by testing the proposed framework on additional crisis datasets, including CrisisLex, CrisisBench, HumAID, CrisisMMD, and other event-specific benchmarks. Future work will also investigate multilingual disaster text classification for crisis monitoring in multilingual and low-resource regions. In addition, real-time Twitter/X stream analysis, deployment-oriented optimization, and domain-informed attention mechanisms incorporating crisis lexicons, temporal signals, geographic indicators, and event-level metadata will be explored.

Acknowledgement: The authors would like to express their sincere gratitude to the Centre for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM), for its valuable support of this research.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Sheraz Ali Hassan and Hamid Masood Khan; methodology, Muhammad Javed and Fida Muhammad Khan; software, Sheraz Ali Hassan and Fida Muhammad Khan; validation, Mohd Faizal Bin Yusof, Jamil Abedalrahim Jamil Alsayaydeh, and Inam Ullah; formal analysis, Hamid Masood Khan and Muhammad Javed; investigation, Sheraz Ali Hassan, Hamid Masood Khan, and Muhammad Javed; resources, Mohd Faizal Bin Yusof and Jamil Abedalrahim Jamil Alsayaydeh; data curation, Fida Muhammad Khan and Muhammad Javed; writing—original draft preparation, Sheraz Ali Hassan, Hamid Masood Khan, and Muhammad Javed; writing—review and editing, Mohd Faizal Bin Yusof, Jamil Abedalrahim Jamil Alsayaydeh, Fida Muhammad Khan, and Inam Ullah; visualization, Sheraz Ali Hassan and Fida Muhammad Khan; supervision, Jamil Abedalrahim Jamil Alsayaydeh and Inam Ullah; project administration, Muhammad Javed, Mohd Faizal Bin Yusof, and Inam Ullah. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: The dataset used in this study was obtained from the "Natural Language Processing with Disaster Tweets" competition hosted on Kaggle. It is publicly available for academic research and can be accessed via the following link: https://www.kaggle.com/competitions/nlp-getting-started/data.

Ethics Approval: Not applicable. This study used a publicly available social media benchmark dataset for aggregate disaster-related tweet classification. No human participants were recruited, no animal subjects were involved, and no private identifiable information was collected by the authors. Therefore, formal ethical approval was not required.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Abboodi B, Pileggi SF, Bharathy G. Social networks in crisis management: a literature review to address the criticality of the challenge. Encyclopedia. 2023;3(3):1157–77. doi:10.3390/encyclopedia3030084.
2. Noor N, Okhai R, Jamal TB, Kapucu N, Ge YG, Hasan S. Social-media-based crisis communication: assessing the engagement of local agencies in Twitter during Hurricane Irma. Int J Inf Manag Data Insights. 2024;4(2):100236. doi:10.1016/j.jjimei.2024.100236.
3. Krystallidou D, Braun S. Risk and crisis communication during COVID-19 in linguistically and culturally diverse communities: a scoping review of the available evidence. In: The languages of COVID-19. London, UK: Routledge; 2022. p. 128–44. doi:10.4324/9781003267843-11.

4. Sakhiyya Z, Saraswati GPD, Anam Z, Azis A. What's in a name? Crisis communication during the COVID-19 pandemic in multilingual Indonesia. Int J Multiling. 2024;21(2):1169–82. doi:10.1080/14790718.2022.2127732.
5. Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev. 2022;55(7):5731–80. doi:10.1007/s10462-022-10144-1.
6. Faisal MR, Nugroho RA, Ramadhani R, Abadi F, Herteno R, Saragih TH. Natural disaster on Twitter: role of feature extraction method of Word2Vec and lexicon based for determining direct eyewitness. Trends Sci. 2021;18(23):680. doi:10.48048/tis.2021.680.
7. Imran M, Castillo C, Diaz F, Vieweg S. Processing social media messages in mass emergency: a survey. ACM Comput Surv. 2015;47(4):67:1–38. doi:10.1145/2771588.
8. Yadav P, Kashyap I, Bhati BS. Contextual ambiguity framework for enhanced sentiment analysis. Tehnički Glas. 2024;18(3):385–93. doi:10.31803/tg-20231227064230.
9. Chandra R, Krishna A. COVID-19 sentiment analysis via deep learning during the rise of novel cases. PLoS One. 2021;16(8):e0255615. doi:10.1371/journal.pone.0255615.
10. Hu X, Chu L, Pei J, Liu W, Bian J. Model complexity of deep learning: a survey. Knowl Inf Syst. 2021;63(10):2585–619. doi:10.1007/s10115-021-01605-0.
11. Parsaeimehr E, Fartash M, Akbari Torkestani J. Improving feature extraction using a hybrid of CNN and LSTM for entity identification. Neural Process Lett. 2023;55(5):5979–94. doi:10.1007/s11063-022-11122-y.
12. Al-Selwi SM, Hassan MF, Abdulkadir SJ, Muneer A. LSTM inefficiency in long-term dependencies regression problems. J Adv Res Appl Sci Eng Technol. 2023;30(3):16–31. doi:10.37934/araset.30.3.1631.
13. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555. 2014. doi:10.48550/arXiv.1412.3555.
14. Tan KL, Lee CP, Lim KM. RoBERTa-GRU: a hybrid deep learning model for enhanced sentiment analysis. Appl Sci. 2023;13(6):3915. doi:10.3390/app13063915.
15. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing; 2014 Oct 25–29; Doha, Qatar. p. 1746–51. doi:10.3115/v1/D14-1181.
16. Horvat M, Gledec G, Leontić F. Hybrid natural language processing model for sentiment analysis during natural crisis. Electronics. 2024;13(10):1991. doi:10.3390/electronics13101991.
17. Cavalin P, Pinhanez CS. Theoretical and empirical advantages of dense-vector to one-hot encoding of intent classes in open-world scenarios. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation; 2024 May 20–25; Torino, Italy. p. 16000–13.
18. Johnson SJ, Murty MR, Navakanth I. A detailed review on word embedding techniques with emphasis on Word2Vec. Multimed Tools Appl. 2024;83(13):37979–8007. doi:10.1007/s11042-023-17007-z.
19. Dharrao D, Aadithyanarayanan MR, Mital R, Vengali A, Pangavhane M, Rajput S, et al. An efficient method for disaster tweets classification using gradient-based optimized convolutional neural networks with BERT embeddings. MethodsX. 2024;13(1):102843. doi:10.1016/j.mex.2024.102843.
20. Liu J, Singhal T, Blessing LT, Wood KL, Lim KH. CrisisBERT: a robust transformer for crisis classification and contextual crisis embedding. In: Proceedings of the 32nd ACM Conference on Hypertext and Social Media; 2021 Aug 30–Sep 2; Virtual. New York, NY, USA: ACM; 2021. p. 133–41. doi:10.1145/3465336.3475117.
21. Le AD. Disaster tweets classification using BERT-based language model. arXiv:2202.00795. 2022. doi:10.48550/arXiv.2202.00795.
22. Nguyen DQ, Vu T, Nguyen AT. 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 Nov 16–20; Virtual. Kerrville, TX, USA: Association for Computational Linguistics; 2020. p. 9–14. doi:10.18653/v1/2020.emnlp-demos.2.
23. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Kerrville, TX, USA: Association for Computational Linguistics; 2019. p. 4171–86. doi:10.18653/v1/N19-1423.

24. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019. doi:10.48550/arXiv.1907.11692.
25. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. 2019. doi:10.48550/arXiv.1910.01108.
26. Wahid JA, Shi L, Gao Y, Yang B, Wei L, Tao Y, et al. Topic2Labels: a framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response. Expert Syst Appl. 2022;195:116562. doi:10.1016/j.2022.116562.
27. Imran M, Mitra P, Castillo C. 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; 2016 May 23–28; Portoroz, Slovenia. p. 1638–43.
28. Zhang Y, Tang W, Ni T. A public opinion propagation model for technological disasters. Sci Rep. 2025;15:7809. doi:10.10380/s41598-025-91244-0.
29. Kamruzzaman M, Kim G. Efficient sentiment analysis: a resource-aware evaluation of feature extraction techniques, ensembling, and deep learning models. In: Proceedings of the 11th International Workshop on Natural Language Processing for Social Media; 2023 Nov 1; Bali, Indonesia. Kerrville, TX, USA: Association for Computational Linguistics; 2023. p. 9–20.
30. Kanungo S, Jain S. Hybrid deep neural network G-LSTM for sentiment analysis on Twitter: a novel approach to disaster management. IngéNierie Syst D'inf. 2023;28(6):1565–75. doi:10.18280/isi.280613.
31. Tan KL, Lee CP, Lim KM, Anbananthen KSM. Sentiment analysis with ensemble hybrid deep learning model. IEEE Access. 2022;10(12):103694–704. doi:10.1109/ACCESS.2022.3210182.
32. Yunida R, Faisal MR, Muliadi I, Abadi F, Budiman F, Prastya I, et al. LSTM and Bi-LSTM models for identifying natural disasters reports from social media. J Electron Electromed Eng Med Inform. 2023;5(4):241–9. doi:10.35882/jeeemi.v5i4.319.
33. Mu G, Liao Z, Li J, Qin N, Yang Z. IPSO-LSTM hybrid model for predicting online public opinion trends in emergencies. PLoS One. 2023;18(10):e0292677. doi:10.1371/journal.pone.0292677.
34. Ahmad Z, Jindal R, Mukuntha NS, Ekbal A, Bhattachharyya P. Multi-modality helps in crisis management: an attention-based deep learning approach of leveraging text for image classification. Expert Syst Appl. 2022;195:116626. doi:10.1016/j.2022.116626.
35. Upadhyay A, Meena YK, Chouhan SS. EJMACC: emotion aided a joint learning approach for multi-aspect crisis event classification. Expert Syst Appl. 2025;282:127597. doi:10.1016/j.2025.127597.
36. Masethe MA, Masethe HD, Ojo SO. Context-aware embedding techniques for addressing meaning conflation deficiency in morphologically rich languages word embedding: a systematic review and meta-analysis. Computers. 2024;13(10):271. doi:10.3390/computers13100271.
37. Lin SC, Lin J. A dense representation framework for lexical and semantic matching. ACM Trans Inf Syst. 2023;41(4):1–29. doi:10.1145/3582426.
38. Olteanu A, Castillo C, Diaz F, Vieweg S. CrisisLex: a lexicon for collecting and filtering microblogged communications in crises. Proc Int AAAI Conf Web Soc Media. 2014;8(1):376–85. doi:10.1609/icwsm.v8i1.14538.
39. Alam F, Qazi U, Imran M, Ofli F. HumAID: human-annotated disaster incidents data from Twitter with deep learning benchmarks. Proc Int AAAI Conf Web Soc Media. 2021;15(1):933–42. doi:10.1609/icwsm.v15i1.18116.
40. Alam F, Sajjad H, Imran M, Ofli F. CrisisBench: benchmarking crisis-related social media datasets for humanitarian information processing. Proc Int AAAI Conf Web Soc Media. 2021;15(1):923–32. doi:10.1609/icwsm.v15i1.18115.
41. Alam F, Ofli F, Imran M. CrisisMMD: multimodal Twitter datasets from natural disasters. Proc Int AAAI Conf Web Soc Media. 2018;12(1):465–73. doi:10.1609/icwsm.v12i1.14983.
42. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2(1):37–63.
43. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Technol. 2006;7(1):1–30.