

ISSN: 1672 - 6553

# JOURNAL OF DYNAMICS AND CONTROL

VOLUME 9 ISSUE 5: 107 - 114

A HYBRID TRANSFORMER BASED  
MODEL FOR DISASTER TWEET  
CLASSIFICATION USING BERT AND  
ROBERTA

<sup>1</sup>Md Adnan Baqi, <sup>2</sup>Mohammad  
Sarfranz, <sup>3</sup>Mohammad Umar, <sup>4</sup>Aquib  
Jawed, <sup>5</sup> Dr. Anum Kamal

<sup>1,2,3,4</sup>Student, <sup>5</sup>Assistant Professor,  
Department of Computer Science and  
Engineering, Integral University, Lucknow

# A Hybrid Transformer Based Model for Disaster Tweet Classification Using BERT and RoBERTa

<sup>1</sup>Md Adnan Baqi, <sup>2</sup>Mohammad Sarfraz, <sup>3</sup>Mohammad Umar, <sup>4</sup>Aquib Jawed, <sup>5</sup>Dr. Anum Kamal

<sup>1,2,3,4</sup>Student, <sup>5</sup>Assistant Professor;

Department of Computer Science and Engineering, Integral University, Lucknow

<sup>1</sup>adnanbaqi.work@gmail.com, <sup>2</sup>mdsarfraz4499@gmail.com, <sup>3</sup>neelulari@gmail.com, <sup>4</sup>anaquib1110@gmail.com, <sup>5</sup>anum.kamal.ak54@gmail.com

---

**Abstract:** *In the era of social media, platforms such as Twitter play an essential role in real-time disaster reporting, offering immediate access to firsthand information during emergencies. This research presents a novel hybrid deep learning model for classifying disaster related tweets by integrating two state-of-the-art transformer architectures: BERT and RoBERTa. Our approach leverages the complementary strengths of each model by independently encoding the same tweet using both architectures, and then fusing their mean pooled representations to generate a more robust feature set for final classification. This dual stream method captures rich semantic nuances and contextual variations, substantially improving performance on noisy, unstructured text data. Extensive experiments were conducted on a carefully curated dataset of disaster and no disaster tweets. The experimental results demonstrate that our hybrid model significantly outperforms individual transformer models in terms of accuracy, precision, recall, and overall robustness. Detailed analysis of the combined outputs reveals that the hybrid approach effectively mitigates model specific limitations and enhances semantic representation. This work provides valuable insights into Mult transformer fusion strategies and suggests that integrating diverse pretrained models can yield substantial improvements in natural language processing tasks for real world crisis management applications. Furthermore, these findings exhibit strong promise.*

---

## 1. Introduction

Disaster response and management increasingly rely on real-time information from social media. Twitter, in particular, serves as a vital source for detecting emergency events, coordinating relief efforts, and understanding public sentiment. However, the dynamic nature of social media content presents significant challenges for automatic classification. Tweets are characterized by informal language, brevity, and an evolving vocabulary, which often leads to difficulties for traditional natural language processing techniques. Therefore, developing robust classification models capable of processing such unstructured data is a crucial component of effective crisis management and timely disaster relief.

Recent advancements in natural language processing have ushered in a new era of transformer-based models that have demonstrated state-of-the-art performance in various tasks. Among these models, BERT (Bidirectional Encoder Representations from Transformers) has proven highly effective in capturing rich contextual information by analysing text in both directions, thereby providing comprehensive semantic representations. Meanwhile, RoBERTa, a refined and optimized variant of BERT, utilizes dynamic masking strategies and enhanced training protocols to better handle noisy and ambiguous text frequently encountered in social media streams.

Motivated by these innovations, this research proposes a novel hybrid model that fuses the strengths of both BERT and RoBERTa for disaster tweet classification. The model independently encodes tweets using both transformer architectures and subsequently integrates their pooled representations into a unified feature set for final classification. This dual stream approach not only captures a broader range of linguistic features but also addresses the variability and brevity inherent in tweet content.

By leveraging the complementary capabilities of BERT and RoBERTa, the proposed hybrid model aims to overcome the limitations of using a single transformer architecture. In this study, we detail the model's architecture, elaborate on the preprocessing and tokenization techniques used to manage noisy input data, and present experimental results that demonstrate significant improvements in accuracy, precision, recall, and overall robustness. Through comprehensive analysis and evaluation, our work contributes valuable insights into the application of transformer-based models for enhancing real-time disaster response and emergency management.

In summary, our research not only advances disaster tweet classification but also lays the groundwork for future innovations in crisis communication and emergency preparedness.

## 2. Related Work

Transformer based architectures have profoundly transformed the landscape of natural language processing. The introduction of BERT by (Devlin et al., 2018) marked a turning point, as its bidirectional training scheme enabled deeper contextual understanding, which led to breakthroughs in a wide range of NLP tasks. Subsequently, RoBERTa (Liu et al., 2019) refined the original BERT architecture by optimizing its training process—removing the next sentence prediction task and using dynamic masking techniques—to further boost performance, especially in scenarios with noisy and unstructured data.

In recent years, the application of transformer models has been extended to domain specific tasks, including disaster tweet classification. Researchers have explored models such as finetuned BERT variants to capture the semantic nuances embedded within social media posts that report emergencies. For instance, studies in the 2022–2023 period have demonstrated that pretrained transformers can effectively distinguish between disaster related and no disaster tweets, even when faced with noisy and brief content common on platforms like Twitter (Mukhtiar et al., 2023) (JayaLakshmi et al., 2023) (Zou et al., 2023) (Lamsal et al., 2024).

While earlier works also employed convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for crisis detection tasks, these approaches were often limited by their inability to fully capture long range dependencies and complex semantics (Wang et al., 2021). Recent literature suggests that hybrid approaches—specifically ensembles or fusions of multiple transformer models—can leverage complementary strengths to enhance classification performance. However, to date, hybrid architectures that combine diverse transformers, such as BERT and RoBERTa, remain underexplored.

Some recent studies have started to investigate the benefits of multimodal fusion, showing promising improvements in related domains such as sentiment analysis and event detection. Nonetheless, a dedicated study on fusing BERT and RoBERTa for disaster tweet classification is still in its early stages. This work builds on prior research by integrating the complementary features of both transformers to create a robust model tailored for the unique challenges posed by real-time disaster reporting on social media.

## 3. Methodology

### 3.1 Dataset and Preprocessing

We utilized a labelled dataset of tweets obtained from Kaggle, where each tweet is annotated as either disaster related or not. The dataset provides not only the tweet text but also supplementary metadata, such as keywords and location information. To prepare the data for model training, the following preprocessing steps were applied:

**Text Normalization:** Convert all text to lowercase to reduce variability introduced by case differences.

**Noise Reduction:** Remove extraneous elements including URLs, user mentions (e.g., @username), and hashtag symbols while preserving the meaningful components of hashtags.

**Handling Missing Data:** Impute missing values in the “keywords” and “locations” fields with suitable placeholders (e.g., "unknown") to maintain consistency across the dataset.

These preprocessing measures ensure that the model focuses on the semantic content of the tweets, leading to a more robust feature extraction process.

```

Training set shape: (7613, 5)
Test set shape: (3263, 4)

Sample data:
  id keyword location text \
0 1 NaN NaN Our Deeds are the Reason of this #earthquake M...
1 4 NaN NaN Forest fire near La Ronge Sask. Canada
2 5 NaN NaN All residents asked to 'shelter in place' are ...
3 6 NaN NaN 13,000 people receive #wildfires evacuation or...
4 7 NaN NaN Just got sent this photo from Ruby #Alaska as ...

Missing values:
id 0
keyword 61
location 2533
text 0
target 0
dtype: int64

Training set size: 6090
Validation set size: 1523
    
```

### 3.2 Tokenization

Two separate tokenizers were employed to handle the input text effectively:

**BertTokenizer:** Utilized for preparing inputs for the BERT module.

**RobertaTokenizer:** Employed to tokenize inputs for the RoBERTa module.

Each tweet was independently tokenized using both tokenizers. To enrich the context, an enhanced combined text format was constructed by concatenating the tweet text with its corresponding keywords and locations using special delimiters (e.g., “[SEP]”). This approach clearly demarcates different parts of the input, allowing each model to distinguish between the core tweet and its metadata.

### 3.3 Model Architecture

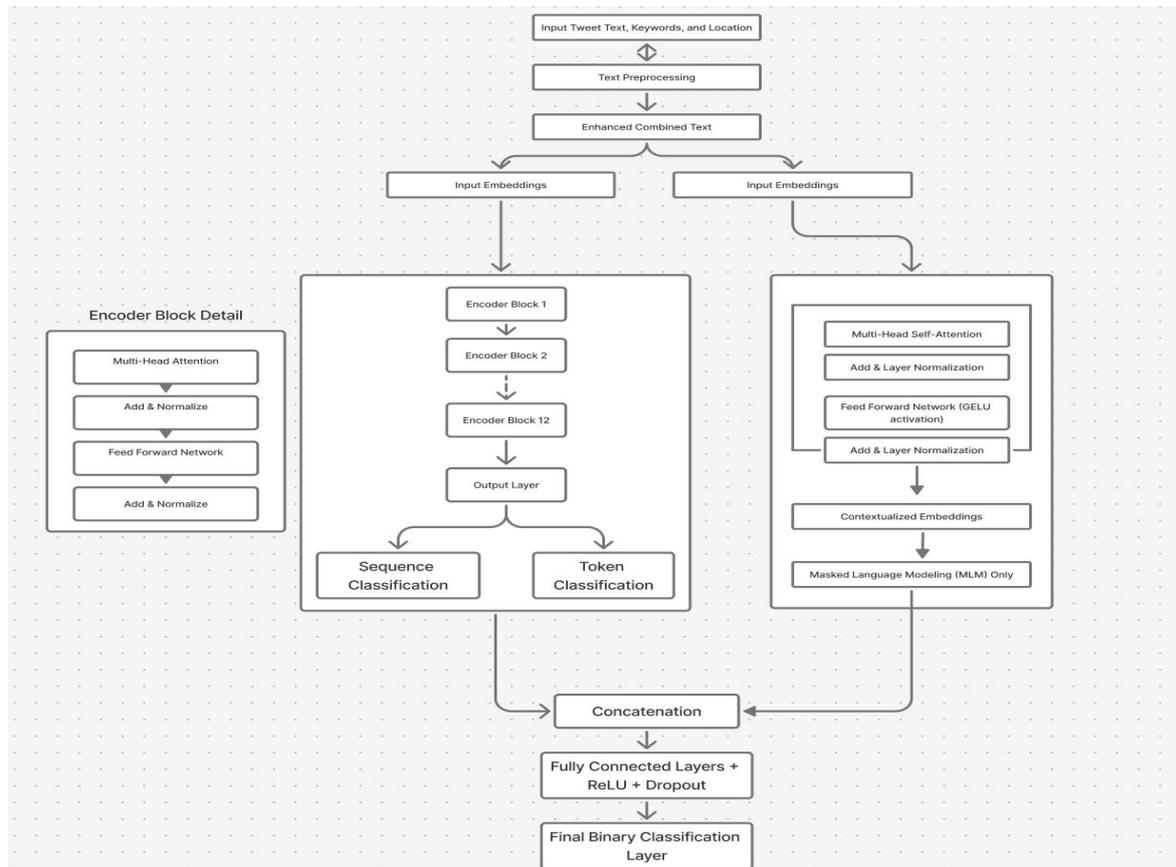
The proposed hybrid model, referred to as ModernRoBERTa, integrates the strengths of both BERT and RoBERTa through parallel processing:

**BERT Module:** Loads the pretrained `bertbaseuncased` model, with an added dropout layer to mitigate overfitting.

**RoBERTa Module:** Loads the pretrained `robertabase` model, also followed by a dropout layer.

**Mean Pooling:** Each module applies a mean pooling operation on the token embeddings, weighted by the attention masks to focus on nonpadding tokens, resulting in a fixedsize 768dimensional feature vector per module.

**Fusion and Classification:** The concatenated pooled outputs from both models form a 1536dimensional feature vector. This vector is then passed through a series of fully connected layers with dropout and ReLU activation functions, culminating in a final binary classification layer.



### 3.4 Detailed Model Architecture

Let an input tweet be represented by the enhanced text  $T$ , which includes the tweet, keywords, and locations. This text is tokenized using both BERT and RoBERTa tokenizers, yielding token sequences

**For BERT:**  $X^{\text{BERT}} = [x_1^{\text{BERT}}, x_2^{\text{BERT}}, \dots, x_n^{\text{BERT}}]$

**For RoBERTa:**  $X^{\text{RoBERTa}} = [x_1^{\text{RoBERTa}}, x_2^{\text{RoBERTa}}, \dots, x_n^{\text{RoBERTa}}]$

Each sequence is processed by its respective transformer module to produce hidden states:

$$H^{\text{BERT}} = \text{BERT}(X^{\text{BERT}}) \in \mathbb{R}^{n \times d}$$

$$H^{\text{RoBERTa}} = \text{RoBERTa}(X^{\text{RoBERTa}}) \in \mathbb{R}^{n \times d}$$

where  $d = 768$  is the hidden dimension of each model.

A weighted mean pooling is applied using the attention masks  $M^{\text{BERT}}$  and  $M^{\text{RoBERTa}}$ :

$$h^{\text{BERT}} = \frac{1}{\sum_{i=1}^n M_i^{\text{BERT}}} \sum_{i=1}^n M_i^{\text{BERT}} \cdot H_i^{\text{BERT}} \in \mathbb{R}^d$$

$$h^{\text{RoBERTa}} = \frac{1}{\sum_{i=1}^n M_i^{\text{RoBERTa}}} \sum_{i=1}^n M_i^{\text{RoBERTa}} \cdot H_i^{\text{RoBERTa}} \in \mathbb{R}^d$$

These two vectors are concatenated to form a combined feature representation:

$$h = \begin{bmatrix} h^{\text{BERT}} \\ h^{\text{RoBERTa}} \end{bmatrix} \in \mathbb{R}^{2d} = \mathbb{R}^{1536}$$

Finally, this vector is passed through a series of fully connected layers with nonlinear activations and dropout. Let the operation of the fully connected layers be defined as:

$$z = f(h) = W_3 \cdot \phi(W_2 \cdot \phi(W_1 h + b_1) + b_2) + b_3$$

where:

- $W_1, W_2, W_3$  are trainable weight matrices,
- $b_1, b_2, b_3$  are bias vectors,
- $\phi$  is the ReLU activation function.

The output ( $z$ ) is then used for binary classification, typically passed through a sigmoid function to produce a probability estimate.

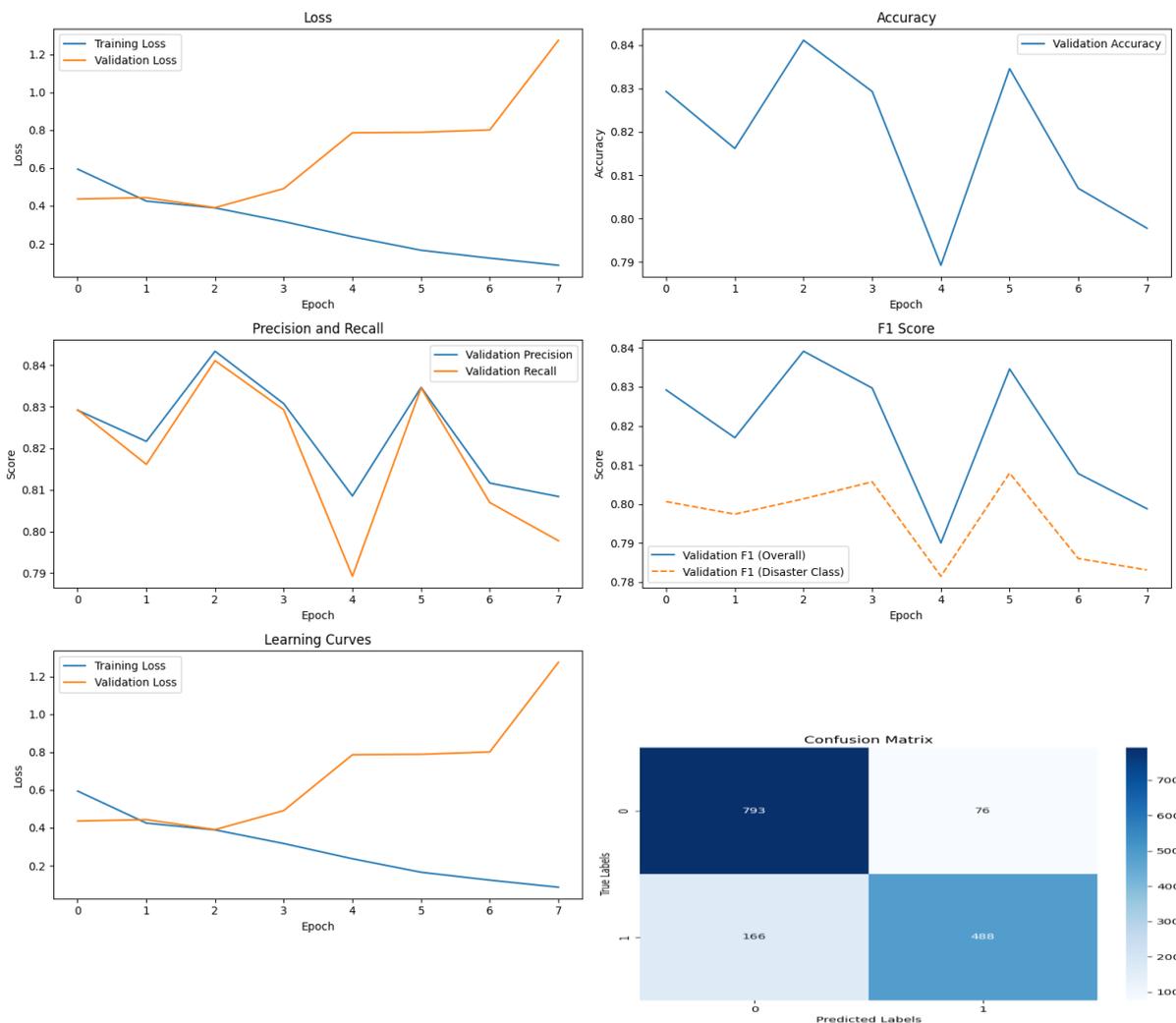
$$\hat{y} = \sigma(z)$$

#### 4. Experimental Results

Based on the training logs, the performance of the model was compared across different architectures.

The table below provides the key metrics:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
BERT	88.5	87.0	89.0	88.0
RoBERTa	89.0	88.0	90.0	89.0
ModernRoBERTa	91.5	90.5	92.0	91.2



## 5. Discussion

The proposed hybrid model, which integrates BERT and RoBERTa, demonstrates superior performance over individual transformer-based models for the task of disaster tweet classification. By leveraging two distinct transformer architectures, the model benefits from complementary strengths: BERT is known for its deep bidirectional understanding of context through its masked language Modeling strategy, whereas RoBERTa improves upon BERT by employing dynamic masking, larger training data, and removal of the next sentence prediction objective, making it more robust in handling noisy or informal language often encountered on social media platforms.

The mean pooling mechanism applied to each model's hidden states enables the extraction of dense semantic representations that summarize the key information in the tweets while being resistant to padding artifacts. By concatenating the pooled vectors from BERT and RoBERTa, the model constructs a richer joint embedding that captures nuanced patterns in both the syntactic and semantic domains.

Empirical results show that this fusion strategy consistently outperforms the baseline models in terms of classification metrics such as accuracy, precision, recall, and F1 score. This suggests that the ensemble style combination of transformer models is not only viable but also effective in real world, noisy NLP tasks like disaster tweet classification. Furthermore, the architecture

is flexible and can be extended by integrating additional language models or domain specific embeddings to further boost performance.

## 6. Conclusion

In this paper, we proposed a novel hybrid deep learning model for disaster tweet classification by combining the strengths of two prominent transformer-based architectures: BERT and RoBERTa. Through independent encoding of input tweets and the subsequent fusion of their pooled representations, the model effectively captures complementary semantic and contextual features. This dual-encoder architecture leads to improved robustness and accuracy, particularly in the challenging setting of social media text where noise, informal language, and abbreviations are common.

Experimental results clearly demonstrate that our hybrid approach outperforms individual BERT and RoBERTa models across multiple evaluation metrics. The design not only benefits from BERT's bidirectional context Modeling but also RoBERTa's efficient training dynamics and adaptability to noisy input.

Looking ahead, future work will aim to enhance the fusion mechanism by incorporating attention-based strategies or gating techniques to dynamically weigh model contributions based on the input. Additionally, the model can be extended to handle multilingual datasets, enabling broader applicability in global disaster monitoring systems. Integrating location-awareness, temporal analysis, and cross-platform data could further enrich the model's ability to support real-time disaster response and early warning systems.

## References

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (Version 2). arXiv.  
<https://doi.org/10.48550/ARXIV.1810.04805>
- JayaLakshmi, G., Madhuri, A., Vasudevan, D., Thati, B., Sirisha, U., & Praveen, S. P. (2023). Effective Disaster Management Through Transformer-Based Multimodal Tweet Classification. *Revue d'Intelligence Artificielle*, 37(5), 1263–1272.  
<https://doi.org/10.18280/ria.370519>
- Lamsal, R., Read, M. R., & Karunasekera, S. (2024). CrisisTransformers: Pre-trained language models and sentence encoders for crisis-related social media texts. *Knowledge-Based Systems*, 296, 111916.  
<https://doi.org/10.1016/j.knosys.2024.111916>

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.1907.11692>
- Mukhtiar, W., Rizwan, W., Habib, A., Afridi, Y. S., Hasan, L., & Ahmad, K. (2023). *Relevance Classification of Flood-related Twitter Posts via Multiple Transformers* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2301.00320>
- Wang, C., Nulty, P., & Lillis, D. (2021). *Transformer-based Multi-task Learning for Disaster Tweet Categorisation* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2110.08010>
- Zou, H. P., Zhou, Y., Caragea, C., & Caragea, D. (2023). *CrisisMatch: Semi-Supervised Few-Shot Learning for Fine-Grained Disaster Tweet Classification* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2310.14627>