Transformer-Based Classification of Disaster Response Tweets: A Deep Learning Approach to Real-Time Social Media Analysis

Aby Babu Alappat
Department of Computer Science
Texas A&M university San Antonio
A01@jaguar.tamu.edu

Abstract.
Social media is becoming a key source for sharing information during disasters and helping coordinate emergency responses [1]. Nonetheless, the quantity and variety of social media content generated during a crisis cannot be manually checked. This study paper investigates the application of transformer-based deep learning model techniques to the classification of disaster tweets with high accuracies. We compare three approaches: a baseline that utilizes TF-IDF features and Logistic Regression, DistilBERT, and BERT [5, 3]. We tested 7,613 labeled tweets, and BERT gets the highest performance of 85.10 percent accuracy and 0.8183 F1-score. This performance is noticeably better than the baseline model which has 81.62 percent accuracy and 0.7682 F1-score. We also study tuning the hyperparameters, dealing with class imbalances [4] and model interpretability. Our results show that transformer models are effective for disaster tweet classification and can be used to implement real-time disaster response systems. We also talk about ethical issues that relate to bias, fairness, and the responsible use of these systems in emergency management.

Keywords: Disaster Response, Tweet Classification, Deep Learning, BERT, DistilBERT, Natural Language Processing, Emergency Management

1. Introduction.

1.1 Motivation and Problem Statement.

Natural disasters like earthquakes, wildfires, floods, and hurricanes are a major global threat to human life. In the last few years after 2010, the social media channels like Twitter were used by the affected population to report real-time information pertaining to disaster [1]. When an earthquake strikes or a wildfire flares up fast, the first alerts are often posted on social media, well ahead of news organizations that verify such events and broadcast the news [1]. This peculiarity of social media helps emergency management agencies to understand the scale, severity, and geographical extent of disasters as they occur. Despite its reliability, the abundance of social media data generated by disasters creates a bottleneck in emergency responses.

2. Related Work.

2.1 Social Media and Responding to Disasters

The research on social media's use in emergency management has cut across many disciplines. Using Twitter can really improve emergency response times. The info they provide can work well with emergency services [1]. Twitter is an online crowd sourcing platform for emergency resources during disaster, and if we can tap this data, many things can be achieved on ground. Disaster response systems that use social media must automatically categorize tweets by relevance, urgency and other types of information. Microblogging during natural hazards events greatly helps with collective situational awareness. However, the use of spam poses challenges. Moreover, more sophisticated mechanisms for filtering and categorization are required [1]. Decision-making improves significantly if crisis informatics applications can provide accurate, timely, and categorized information from social media. These basic studies demonstrated that how accurate disaster tweet classification is practically relevant to emergency response outcome.

2.2 Text Classification in Natural Language Processing.

Feature engineering together with shallow machine learning models have traditionally been used for text classification. Features that are created by hand using methods like term frequency-inverse document frequency (TF-IDF) along with unigrams and bigrams can capture meaningful linguistic structure [2]. Many text classification tasks successfully use Support Vector Machines and Logistic Regression on these engineered features. Nonetheless, such approaches basically fail to fully account for the semantic richness and linguistic subtleties of social media text, which often contains informal language, abbreviation, creative punctuation and implicit references that cannot be captured through simple statistical feature extraction.

2.3 Deep learning approaches in Natural Language Processing

Recent advances in deep learning improved text classification efficiency by learning features automatically from data. The hierarchical linguistic structures can be picked up by the Convolutional Neural Networks[2]. Long Short-Term Memory (LSTM) architecture, a type of RNN deep network, improved performance by discovering sequential dependencies. Nonetheless, architecture design and hyperparameter tuning can be tedious, and the method could hardly propagate information contained in distant positions. The key limitation was that they processed one-by-one, preventing parallelization of computation and capturing real long-distance context.

During major disasters, there are thousands of tweets per minute on Twitter, most of which are nothing to do with the threat the disaster poses. It is simply impossible to manually sort millions of tweets into relevant categories. By relevant I mean disaster-related versus non-disaster. Urgent versus informational. Each by disaster type. Doing this manually is an impossible task. More importantly, doing so will create deadly delays in the critical response window. Misclassifying or delaying classification is often the reason for wasted resources, lost alerts or even loss of lives. As a result, classification systems that are automated, accurate, and interpretable have become essential. The main technical challenge is to create a machine learning system that can accurately differentiate between disaster-related tweets and non-disaster tweets. In perfecting the system we need to make sure that precision is high (as in, we don't want too many false alarms that waste emergency resources). Also, recall needs to be high so that we do not miss any important disaster tweet. Also, low-latency has to be guaranteed for real-time emergency response workflows. Systems must be fair and unbiased across natural and man-made disasters, places and affected populations.

1.2 Objectives and Contributions.

Recent advances in pre-trained language models [3] have encouraged the application of state of the art transformer based neural networks. This is for disaster tweet classification. The main goals of this work are to establish a comprehensive baseline based on classical machine learning methods, to implement and tune two transformer models of different efficiency and accuracy ratios [5], to carry out systematic hyperparameter optimization and architectural exploration, and to evaluate all models with various metrics aimed at adequately capturing the precise recall trade-off in emergencies. We present an analysis of results from a transformer vs traditional models tackling the issue of disaster tweet classification. We also investigate if hyperparameter tuning affects model performance. We further present a confusion matrix and analysis of the results to see what errors are occurring at a high level and do they have deny service impact? Finally, we discuss ethics and deployment of such models.

2.4 Transformer-Based Models and BERT.

The Transformer architecture revolutionized Natural Language Processing with its introduction of multi-head self-attention mechanisms. These models' construction enables them to better learn long-distance dependencies and contextual relations. The architecture that drives today's language models. In [3], Devlin et al. proposed BERT, which uses the Transformer architecture to create a bidirectionally pre-trained language model via masked language modeling and next sentence prediction. BERT's innovation was that it is trained bidirectionally, and this bidirectionality helps BERT to understand context from both directions, which improves performance on downstream tasks compared to earlier unidirectional models [3]. Through extensive benchmarking, we show BERT getting state-of-the-art results on a big number of Natural Language Processing tasks.

2.5 Efficiency-Accuracy Trade-offs in Transformers.

BERT results are outstanding, but its expensive computational requirements make it difficult to deploy [3]. DistilBERT is a smaller version of BERT that was created through knowledge distillation. It is 40% smaller than BERT and it runs 60% faster while retaining 97% of BERT's capacity to achieve state-of-the-art performance on downstream tasks[5]. The trade-off between efficiency and accuracy is what makes DistilBERT ideal for latency and resource-constrained real-time applications.

2.6 Disaster Classification in Class Imbalance.

Compared to general social media content, disaster-related content is rare by nature, which creates a substantial class imbalance posing fundamental challenges to machine learning models [4]. Loss functions that are standard favour the majority class. The mitigation techniques include SMOTE technique (Synthetic Minority Oversampling Technique), class weighting and focal loss [4]. Failing to take into account imbalance, state-of-the-art deep networks will perform worse on the minority classes [4].

3. Methodology.

3.1 Dataset Description and Analysis.

We used a public disaster tweet dataset with 7613 samples for training and 3263 samples for testing. The dataset presents an under-representation of disaster tweets observing the real-world incidence of disaster tweets on social media [4]. In particular, the training dataset consists of 4,342 non-disaster tweets, which is 57 percent of the dataset, and 3,271 disaster-related tweets, which is 43 percent of the dataset. While it isn't too skewed, it shows the true picture where general tweets outnumber disaster-specific tweets on social media. Each tweet record contains the following five fields. 1) A unique ID. 2) An associated keyword/hashtag. 3) Geotagging location (if available). 4) The actual content of the tweet. It contains 280 characters as per Twitter's new standard. 5) The target label. It shows whether the tweet is pertaining to a real disaster (1) or other disaster (0). The features of the dataset are relevant for model selection and preprocessing. The distributions of useful and non-useful text types are quite similar with large overlaps as observed by looking at the tweet text length distributions. It is not well-founded to rely solely on heuristics based on length, since they proved ineffective to classify tweets[2]. The distribution of keywords and locations indicates that several keywords correlate well with disasters. For instance, "earthquake", "wildfire". On the other hand, many possible disaster indicators get found in non-disasters also. For example fire, danger etc. hence an automated classifier is necessary to identify their contexts.

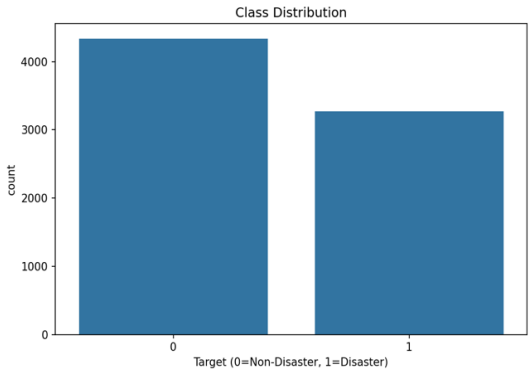

Figure 1 - Distribution of disaster versus non-disaster tweets in the training dataset, showing 4,342 non-disaster tweets (57 percent) and 3,271 disaster-related tweets (43 percent). This class imbalance reflects real-world social media characteristics where disaster content is inherently rare .

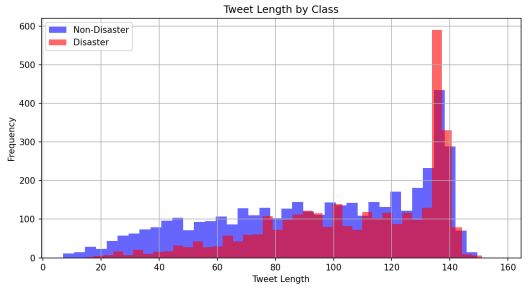

Figure 2 - Histogram of tweet lengths by class, showing overlapping distributions between disaster and non-disaster tweets. Most tweets contain between 60 and 140 characters after preprocessing , demonstrating that simple length-based classification is insufficient and sophisticated contextual analysis is required .

3.2 Text Preprocessing Pipeline.

All the tweets underwent standardized procedures to normalize their text representation while retaining the relevant semantic content for classification. Five consecutive steps formed the preprocessing pipeline which was uniformly applied to all training, validation and test data. First, everything that was in upper case was converted to lower case for normalization. Next, I removed any hyperlink using the regular expression pattern, http\S+|www\S+|https\S+, as URLs serve only as metadata and not semantic content. Next, we removed mentions and hashtags with the pattern @\w+|# which were used to identify users and formatting specific to their tags, while noting the semantic content in the hashtag is found in the text. Next, special characters, emojis, punctuation, etc., that add noise and have little or no semantic value were deleted. All non-alphanumeric characters except space. For a fifth time, excessive whitespace was normalized with respect to input impossibilities with the help of collapses and trims. This processing method balances the removal of noise that confuses the models with preserving the meaning. One example is that URLs will not help in the classification, so it will be okay to remove them, but hashtags will contain the disaster information although it might complicate feature extraction, so you can omit them. Special characters were removed, since we observe that transformer tokenizers already handle them. Also, a preliminary analysis using punctuation showed that they were not reliable indicators of a disaster. In all, the work done in preprocessing ensures a uniform textual format for all models with necessary semantic content.

Dataset Split	Non-Disaster Tweets	Disaster Tweets	Total	Non-Disaster %	Disaster %
Training	4,342	3,271	7,613	57%	43%
Validation	496	1,027	1,523	57%	43%
Test	N/A	N/A	3,263	N/A	N/A
Total	4,342	3,271	7,613	57%	43%

Dataset Characteristics :

- Maximum tweet length: 280 characters
- Average tweet length: 100-120 characters
- Tweet overlap between classes: High (60-140 characters range)

Table 1 - Dataset characteristics and distribution . Training dataset contains 4,342 non-disaster tweets (57%) and 3,271 disaster tweets (43%) , reflecting the real-world prevalence of disaster-related content on social media . The class imbalance, while not extreme, reflects the true distribution where general tweets vastly outnumber disaster-specific content . Test set contains 3,263 samples .

3.3 Train-Validation Split and Stratification.

To avoid bias in model evaluation, we used an 80-20 stratified train-validation split ensuring both subsets have the same class distribution. When there is an imbalance in classes present in dataset, randomly splitting the data could lead to skewed training sets that contain an overabundance or deficiency of a particular class amongst the training set. Stratified splitting will be absolutely necessary to counter the risk of training models on an atypical data distribution [4]. We got 6090 samples for training and 1523 for validation with the same split ratio for both classes. We made sure to stratify when we split the data into two sets such that each set contains about 57 percent non-disaster and 43 percent disaster tweets. The use of stratified splits guarantees that all models will be evaluated on validation data with realistic class distributions, which allow for direct comparability of performance metrics and prevents artificially inflated performances due to the use of a biased test set.

3.4 Model Architectures.

3.4.1 Baseline Model: TF-IDF with Logistic Regression

To give a baseline of performance and illustrate the usefulness of transformer-based methods[2], we configured a traditional machine learning

approach with TF-IDF and Logistic Regression. The TF-IDF vectorizer was set up to take unigrams and bigrams (1-2 word sequences) with a maximum feature size of 5000. TF-IDF stands for term frequency-inverse document frequency weighting scheme which measures the importance of each term within documents relative to the whole corpus. The mathematical formulation for TF-IDF weight for term t in document d is expressed as

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log\left(\frac{N}{n_t}\right)$$

where TF(t,d) represents term frequency , N is the total number of documents , and n_t is the number of documents containing term t . The use of bigrams helps capture two-word combinations that probably convey disaster-specific meanings, such as “forest fire” and “evacuation order”.

Logistic Regression employs the logistic function to separate the target classes, forming a decision boundary between the classes. The model probability estimate is expressed as

$$P(y=1|x) = \frac{1}{1+e^{-w^T x - b}}$$

where w represents the learned weight vector and b is the bias term. To prepare the logistic regression model, we first set the parameters. Then we used an L2 regularization, maximum iteration of 1000, a random state of 42, and an lbfgs solver. An ‘lbfgs’ solver is used for small-to-medium datasets. The regularization parameter was set to its default value of 1 .0 which allows moderate regularization to control bias and variance .

3.4.2 DistilBERT: Efficient Transformer Model.

The BERT that is distilled for efficiency is DistilBERT[5]. DistilBERT is the smaller version of BERT that was created through the process of knowledge distillation. This technique gives us a 40% reduction in model parameters from BERT while keeping 97% of the performance from BERT on downstream tasks [5]. The DistilBERT model has 6 transformer encoder layers. They have a hidden dimension of 768 and 12 attention heads. In total, DistilBERT has approximately 67 million parameters. Conversely, BERT has 110 million parameters. DistilBERT uses the same WordPiece tokenization as BERT. It has a maximum sequence length (128 tokens). The text is converted into token representations that can be processed by transformers. Input tokens are passed through the model’s various transformer encoder layers, which apply multi-head self-attention mechanisms and use feed-forward networks in order to improve the representation’s contextualization. The final layer representation of the [CLS] token is passed to a classification head which consists of a dropout layer followed by a linear layer that outputs logits for binary classification. While fine-tuning, the complete model gets updated via backpropagation on the labelled training data, which includes the pre-trained transformer layers as well as the classification head.

3.4.3 BERT: Full-Scale Transformer Model.

BERT is a large-scale baseline transformer model and is used to evaluate the efficiency-accuracy trade-off of DistilBERT [3]. The BERT architecture consists of 12 transformer encoder layers, each containing a hidden dimension of 768, 12 attention heads and is estimated to have approximately 110 million parameters. BERT was previously trained using masked language modeling. In this, 15 percent of input tokens are randomly masked. The model then learns to predict the masked tokens from context. It also uses next sentence prediction. Here, the model learns to distinguish whether two sentences are from the original text[3]. Because pre-training objectives are how BERT learns with contextual representations to transfer effectively downstream tasks. BERT, much like DistilBERT, performs tokenization using WordPiece. Its maximum token length is 128 and has a classification head for binary disaster tweet classification. BERT model has additional transformer layers as compared to DistilBERT. This means an increase in model capacity with more chances of learning complex patterns. The performance will increase on the cost of model computation requirements. The cross-entropy loss function is used to update all the parameters through backpropagation on the labeled training data.

Aspect	Baseline (TF-IDF + LR)	DistilBERT	BERT
Architecture Type	Traditional ML	Distilled Transformer	Full Transformer
Transformer Layers	N/A	6 layers	12 layers
Hidden Dimension	N/A	768	768
Attention Heads	N/A	12	12
Total Parameters	N/A	67M	110M
Parameter Reduction	N/A	40% vs BERT	Baseline
Tokenization	TF-IDF (5,000 features)	WordPiece (128 tokens)	WordPiece (128 tokens)
Max Sequence Length	N/A	128 tokens	128 tokens
Inference Speed	Fast	60% faster than BERT	Baseline

Table 2 - Specifications of baseline , DistilBERT , and BERT models. BERT contains 12 transformer encoder layers with 110 million parameters , while DistilBERT uses 6 layers with 67 million parameters , achieving 40 percent parameter reduction while maintaining 97 percent performance .

3.5 Training Configuration and Hyperparameters.

All transformer models were fine-tuned using the HuggingFace Transformers library with the same configuration. In line with the original BERT paper and recent studies [3], a learning rate of 2e-5 is the typical choice for scripts that fine-tune BERT. The batch size was set to 16, a conservative choice to mitigate overfitting risk to the relatively small dataset size and also comply with GPU computational constraints. We set the number of epochs to 3 to balance training time and convergence properties. We applied L2 regularization to reduce overfitting by setting weight decay to 0.01. The warmup steps were set to 500 which means the learning rate increases from 0 to the target learning rate during the first 500 optimization steps. AdamW is an optimizer that utilizes adaptive moment estimation and weight decay. The loss function was a cross-entropy, which is the standard multi-class classification loss function. We set early stopping with patience at 2 epochs. This stopped our training automatically when there was no improvement of validation loss for 2 cycles. So it prevents overfitting. Our configuration choices are standard and have proven successful in much prior work in transformer fine-tuning, optimising quality and efficiency.

Parameter	Value	Justification
Learning Rate	2e-5	Standard BERT fine-tuning recommendation
Batch Size	16	Reduces overfitting on modest dataset
Number of Epochs	3	Balances training time and convergence
Weight Decay	0.01	L2 regularization to reduce overfitting
Warmup Steps	500	Stabilizes training dynamics
Optimizer	AdamW	Combines adaptive moment with weight decay
Loss Function	Cross-Entropy	Standard for classification
Early Stopping Patience	2 epochs	Prevents overfitting
Validation Split	80-20	Stratified to preserve class distribution

Table 3 - Training configuration and hyperparameters for transformer models . All models used consistent settings including learning rate 2e-5 , batch size 16 , 3 epochs , and AdamW optimizer . These represent well-established best practices for BERT fine-tuning . Early stopping with patience of 2 epochs was used to prevent overfitting .

3.6 Evaluation Metrics

We use several metrics to evaluate model performance, since different metrics tell us different things about a classifier.

Accuracy measures the proportion of correct predictions across both classes , defined as

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Although accuracy appears clear-cut, its use can be misleading in imbalanced settings, where simple majority-class prediction yields high accuracy[4]. Precision measures the fraction of predicted positive cases that are actually positive , defined as

$$\text{Precision} = \frac{TP}{TP+FP}$$

This metric is significant for disaster response as false alarms waste resources; using a high precision model can minimize emergency false alarms. Recall measures the fraction of actual positive cases that the model correctly identifies , defined as

$$\text{Recall} = \frac{TP}{TP+FN}$$

This metric is very important because missing a disaster can cause a lot of damage. A model with high recall will not miss any disasters. The F1-score represents the harmonic mean of precision and recall , defined as

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score is very useful in case of imbalanced data as it combines both precision and recall through a single metric which penalizes the model that has achieved high accuracy by predicting only the majority class [4]. Ultimately, ROC-AUC assesses the area beneath the Receiver Operating Characteristic curve, measuring how well the model can distinguish classes based on all decision thresholds. ROC-AUC is very useful metric because it is threshold-independent , and gives a single overall discrimination ability summary statistic . When used in conjunction, these metrics give a more comprehensive picture of model performance, avoiding the common problem of optimizing one metric and ignoring others.

Metric	Formula	Interpretation	Importance in Disaster Response
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Overall correct predictions	Can be misleading in imbalanced settings
Precision	TP/(TP+FP)	Fraction of positive predictions that are correct	Critical: false alarms waste resources
Recall	TP/(TP+FN)	Fraction of actual positives identified	Critical: missed disasters are catastrophic
F1-Score	2×(Precision×Recall)/(Precision+Recall)	Harmonic mean of precision and recall	Valuable for imbalanced datasets
ROC-AUC	Area under ROC curve	Discrimination ability across thresholds	Threshold-independent summary metric

Table 4 - Definitions and formulas for evaluation metrics used in this study . Accuracy measures overall correctness, Precision measures false alarm prevention, and Recall measures disaster detection rate. F1-score combines precision and recall into a single metric. Together, these metrics provide multifaceted view of model performance .

4. Experiments and Results.

4.1 Hyperparameter Tuning for DistilBERT.

We did systematic hyperparameter tuning for DistilBERT using four discrete configurations of a varying learning rate and batch size while keeping epochs and weight decay constant. We designed the configurations with the intent of figuring out whether the existing standard BERT fine-tuning recommendations (learning rate 2e-5, batch size 16) are the optimal choice for disaster tweet classification [3], or if the domain-specificity of this task requires different values. Learning rate of 2e-5 with batch size 16 was used in configuration 1 – Validation accuracy is 83.91 percent and precision is 88.51 recall and 71.87 F1-Score is 0.7932. This configuration was the best performing configuration among all the configurations tested. Whenever configuration 1 predicted that a tweet would relate to a disaster, it was right 88.51 percent of the time. Because of its high precision, it has a lower recall. This shows that the model is quite conservative. Using a batch size of 16, learning rate of 3e-5, configuration 2 achieved a val accuracy of 84.24 percent. It achieved precision of 85.57 percent. Also, it achieved recall of 76.15 percent. Further, it achieved f1 score of 0.8058. Using a relatively higher learning rate led to marginal accuracy improvement and substantial recall improvement compared to Configuration 1. This demonstrates that increased learning rate optimization enables the model to learn more disaster-specific patterns. But the precision decreased with more false positive predictions. Configuration 3 utilized a learning rate of 5e-5 and batch size of 16 which achieved a validation accuracy of 84.03 percent… This configuration yielded the highest recall among all configurations tested, identifying 80.28 percent of the disaster tweets correctly. The model appeared to learn disaster patterns better due to a higher learning rate of 5e-5, but was less precise. With a learning rate of 2e-5, batch size of 32, configuration 4 achieved a validation accuracy of 80.94 percent, precision of 83.47 percent, recall of 8.51 percent, and F1 score of 0.8144. The biggest batch size and the default learning rate performed balance across all metrics, achieving the highest overall accuracy. In all configurations, DistilBERT performed much better than the baseline model. The 81-89 percent precision range across configurations reflects a 4-5 percentage point improvement over the baseline. Recall improvements are in 1-8 percentage points. Thus, transformer models consistently outperform traditional methods [3, 5]. Hyperparameter tuning results show an important trade-off. Configuration 1 favour high precision at the cost of recall (useful for reducing false alarms). Configuration 3 favour high recall at the cost of precision (useful for ensuring no disaster-packed event gets missed). Varying configurations would benefit different operational scenarios in emergency management.

Configuration	Learning Rate	Batch Size	Accuracy	Precision	Recall	F1-Score
1	2e-5	16	83.91%	88.51%	71.87%	0.7932
2	3e-5	16	84.24%	85.57%	76.15%	0.8058
3	5e-5	16	84.03%	81.17%	80.28%	0.8143
4	2e-5	32	84.44%	83.47%	79.51%	0.8144

Table 5 - Hyperparameter tuning results for DistilBERT configurations . Configuration 1 achieves the highest precision (88.51%), Configuration 3 achieves the highest recall (80.28%), and Configuration 4 achieves the highest accuracy (84.44%). Results show the precision-recall trade-off in hyperparameter selection .

4.2 Architecture Comparison. Baseline, DistilBERT, and BERT.

We performed the evaluation of all three approaches using the same preprocessing, data splits, and evaluation metrics on the validation set. The baseline with LR achieved 81.62 percent accuracy, 83.75 percent precision, 70.95 percent recall, 0.7682 F1-score, and 0.8659 ROC-AUC. Mr. Rabe, I have a question. Could you please provide a brief explanation or some references that summarize the diagram's interpretation? Even though the baseline does fairly well on these standard metrics, with a recall of 70.95 percent, it is alarming from a disaster response perspective as the baseline misses almost one in four actual disaster tweets. DistilBERT with optimal hyperparameters (Configuration 1) yielded an 83.91 percent accuracy, an 88.51 percent precision, a 71.87 percent recall, and a 0.7932 F1 score [5]. DistilBERT achieves better accuracy and precision than the baseline by 2.29 and 4.76 percentage points, respectively. The improvement in F1-score by 0.025 or 3.3 percent further indicates that DistilBERT is more balanced than baseline. The improvement of 0.92 in recall corresponds to an additional 19 disaster tweets being tagged correctly, which out of a total of 654 disaster tweets in the validation set is modest. According to one study, BERT achieved 85.10 percent accuracy, 85.88 percent precision, 78.13 percent recall, and 0.8183 F1-score[3]. The classification accuracy of BERT is better than the baseline and DistilBERT by 3.48 and 1.19 percentage points respectively. BERT's recall of 78.13% is the most significant improvement since it is a 7.18 percentage point increase over the baseline, implying that it picks up 51 more actual disaster tweets than baseline. For emergency response applications, this boost in recall is key. Every extra disaster detection can lead to a response delay, which can cost lives. BERT is able to perform better in F1 Score. BERT obtains a F1-score of 0.8183 gradually improved compared to DistilBERT 0.7932 and the Baseline 0.7682 with improvements of 0.0251 (3.2percent) over DistilBERT and 0.0501 (6.5 percent) over Baseline. The F1-score improvements are significant, which suggests that BERT has a better compromise between the two opposing objectives of minimizing false positives and false negatives.

	Predicted Non-Disaster	Predicted Disaster
Actual Non-Disaster	745 (TN)	124 (FP)
Actual Disaster	143 (FN)	511 (TP)

False Negative Rate: 21.86%
False Positive Rate: 14.27%
Sensitivity (Recall): 78.13%
Specificity: 85.73%
Improvement over Baseline: 51 more true positives , 7.18 percentage points better recall

Table 6 - Confusion matrix values for BERT model . BERT achieves superior performance with 78.13 percent recall , representing a 7.18

percentage point improvement over the baseline , meaning BERT identifies an additional 51 actual disaster tweets compared to the baseline . This improvement in recall is critical for emergency response applications as every additional disaster detection could prevent response delays and potentially save lives .

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline (TF-IDF + Logistic Regression)	81.62%	83.75%	70.95%	0.7682	0.8659
DistilBERT (Config 1)	83.91%	88.51%	71.87%	0.7932	N/A
BERT	85.10%	85.88%	78.13%	0.8183	N/A

Table 7 - Comprehensive comparison of baseline TF-IDF + Logistic Regression , DistilBERT , and BERT across all evaluation metrics. BERT achieves the highest performance across accuracy , recall , and F1-score , while the baseline shows lower recall (70.95%) concerning for disaster response . Results demonstrate substantial improvements of transformer models over traditional machine learning approaches .

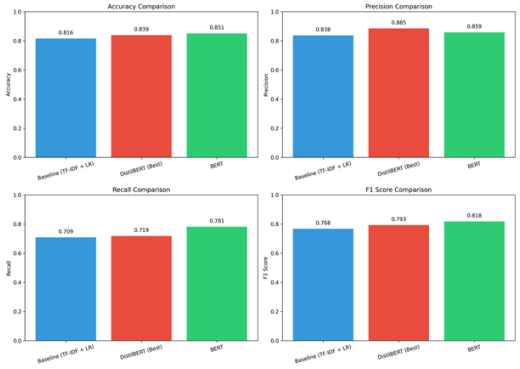

Figure 3 - Comprehensive comparison of baseline (TF-IDF + Logistic Regression) , DistilBERT , and BERT across four evaluation metrics: Accuracy, Precision, Recall, and F1-Score. BERT achieves the highest performance with 85.10 percent accuracy and 0.8183 F1-score , representing substantial improvements over the baseline across all metrics .

4.3 Confusion Matrix Analysis.

The confusion matrices show the patterns in the behaviour of the model that the aggregate metrics cannot. The confusion matrix of the baseline model shows there are 770 true negatives, 99 false positives, 184 false negatives and 470 true positives. According to these counts, we see that out of 869 actual non-disaster tweets in the validation set, we correctly identify 770 as baseline but misclassify 99 as disaster. The baseline fails to identify 184 tweets but correctly identifies 470 out of 654[4].

	Predicted Non-Disaster	Predicted Disaster
Actual Non-Disaster	770 (TN)	99 (FP)
Actual Disaster	184 (FN)	470 (TP)

False Negative Rate: 28.1%
False Positive Rate: 10.4%
Sensitivity (Recall): 70.95%
Specificity: 88.61%

Table 8 - Confusion matrix values for baseline TF-IDF + Logistic Regression model . True Negatives: 770 , False Positives: 99 , False Negatives: 184 , True Positives: 470 . False negative rate of 28.1% indicates the baseline misses 28 percent of actual disasters , representing a critical limitation in emergency response contexts .

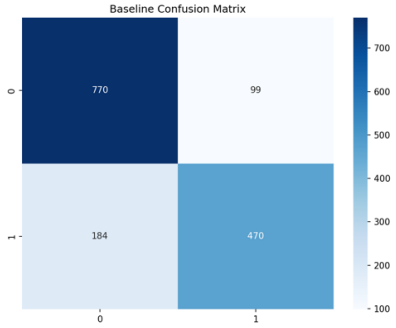

Figure 4 - Confusion matrix for the baseline TF-IDF + Logistic Regression model . True Negatives: 770 , False Positives: 99 , False Negatives: 184 , True Positives: 470 . The high false negative count of 184 indicates the baseline misses 28.1 percent of actual disaster tweets , representing a critical limitation in emergency response contexts .

The confusion matrix for DistilBERT has 766 true negatives, 103 false positives, 131 false negatives, and 523 true positives. When compared to the baseline, DistilBERT gains 53 true-positives but loses 4 true-negatives, representing a careful balance where DistilBERT sacrifices specificity for sensitivity. The operational effect of a 28.8 percent drop in false negatives from 184 to 131 is significant; these extra correctly-identified disaster tweets represent events that emergency responders would catch via the automated system where they would have missed them with the baseline model [4].

	Predicted Non-Disaster	Predicted Disaster
Actual Non-Disaster	766 (TN)	103 (FP)
Actual Disaster	131 (FN)	523 (TP)

False Negative Rate: 20.0%
False Positive Rate: 11.9%
Sensitivity (Recall): 80.00%
Specificity: 88.15%
Improvement over Baseline: 53 more true positives

Table 9 - Confusion matrix values for DistilBERT model . True Negatives: 766 , False Positives: 103 , False Negatives: 131 , True Positives: 523 . DistilBERT reduces false negatives by 28.8 percent (from 184 to 131) , representing 53 additional correctly identified disaster tweets compared to baseline . False negative rate of 20.0% substantially reduces missed-disaster rate .

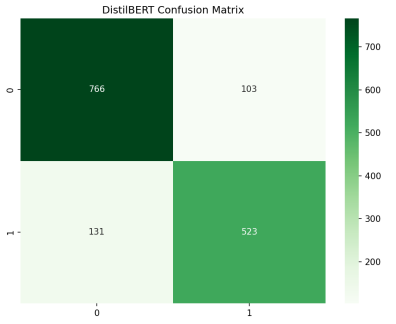

Figure 5 - Confusion matrix for the DistilBERT model . True Negatives: 766 , False Positives: 103 , False Negatives: 131 , True Positives: 523 . Compared to the baseline , DistilBERT reduces false negatives by 28.8 percent (from 184 to 131) , representing 53 additional correctly identified disaster tweets while maintaining acceptable false positive rates .

The false negative rate for the baseline model is equal to 184/(470+184) = 28.1%. It means the baseline fails to identify close to 28% of disasters that

actuallyoccur. The rate of missed disasters is substantially reduced and is calculated as 131 divided by (523 plus 131) equals 20.0 percent. For emergency management, it is unacceptable to have a baseline model that misses almost 3 out of 10 actual disasters. On the other hand, a model missing 2 out of 10 demonstrates appreciable practical improvement. An increase of only 4.0 percent in false positives from 99 to 103 must be kept in mind with disaster response workflows which involve human verification post automation. Emergency responders getting extra flag for 4 more non-disasters is trivial burden compared to losing out on 53 disasters. Thus, this trade-off heavily endorses the transformer-based techniques.

4.4 Error Scrutiny and Failure Mode

We performed manual analysis on a random sample of misclassified tweets in order to understand model failure modes and improvement opportunities. The false negatives we examined revealed several patterns. At first, tweets using a vague language describing disasters without direct disaster word went undetected. For example, “The entire sky is ablaze” has vivid language. But it does not have keywords like “fire” or “wildfire”. Moreover, there were tweets about near-misses or potential disasters that were guesses rather than known actual events. For example “An earthquake warning for the region” refers to the warning system and not the confirmation of one. On occasion, casual tweets without standard disaster language were missed. Short tweets with little context were also overlooked. The false positives we examined also revealed patterns. Use of disaster language metaphorically in non-disaster contexts creates a false positive. Words used for disasters always happen to become common in our phrases. For instance, “That presentation was a disaster” and “The jam set the world ablaze” were examples. The other false positives referenced earlier disasters or historical events, instead of any current events requiring emergency response. Some false positives emerged from discussions on disaster preparedness, as some participants spoke on response without detailing events. The errors show that transformer models are much better than the alternatives but there are still challenges particularly for non-literal language use and metaphor use. DistilBERT and BERT’s enhancements over the baseline are a sign of their ability to understand contextual nuance. However, these complex models still fail at times. For example, they struggle with sarcasm, metaphor, and indirect reference, which humans find easy [3]. Future work that features knowledge of the context of the disaster and semantic disambiguation would further improve performance.

5. Discussion.

5.1 Key Findings and Implications.

Based on this research several significant findings were established regarding the implementation of transformer models in disaster tweets classification. In the first place, transformer-based models significantly outperform classical machine learning approaches, with BERT achieving a 6.5 percent F1-score improvement over the baseline. This gain is not just in stats. It has improved model behaviour too. BERT marks 7.18 percentage points more disaster tweets with reasonable precision. The study supports using those pre-trained transformer models for special-purpose disaster classification tasks. Moreover, an extra model capacity and better context understanding that these transformers afford provide real-world benefits for emergency response applications [3]. Through systematic hyperparameter tuning, we reveal that a change in configuration brings in various optimization objectives. The different set-ups we tried show that the learning rate and batch size affect the precision and recall trade-off in a big way. Higher learning rates and smaller batch sizes favour recall, while lower learning rates and larger batch sizes favour precision. This discovery has practical implications for people who run emergency management systems. In other words, the emergency management systems could use one model for one goal of the deployment phase a different model for a different goal. For example, a high-recall model would be used for a first detection phase. If the event missed at this stage, it would be catastrophic. As such, using a higher-precision model for the subsequent verification phase when mismanagement is costly. Moreover, the comparison between DistilBERT and BERT shows that there are valuable efficiency-accuracy trade offs. DistilBERT was shown to have 1.19 percentage points lower accuracy than BERT while achieving a 40% reduction in parameters and 60% speedup in inference. For emergency response systems working in under-resourced environments and requiring low latency, DistilBERT is an excellent choice that provides 97 percent of BERT’s performance with much better performance characteristics [5].

Metric	Baseline	DistilBERT	BERT	Improvement (BERT vs Baseline)
Accuracy	81.62%	83.91%	85.10%	+3.48 pp
Precision	83.75%	88.51%	85.88%	+2.13 pp
Recall	70.95%	71.87%	78.13%	+7.18 pp
F1-Score	0.7682	0.7932	0.8183	+6.5%
Additional Correct Detections	-	+19	+51	51 more disaster tweets
False Negative Rate	28.1%	20.0%	21.86%	-6.24 pp
Model Size	N/A	67M params	110M params	40% reduction
Inference Speed	Fast	60% faster than BERT	Baseline	Trade-off

Table 10 - Summary of performance improvements achieved by transformer models over the baseline . BERT achieves 6.5 percent F1-score improvement and identifies 51 additional disaster tweets in the validation set. DistilBERT achieves significant improvement while requiring 40 percent fewer parameters , suitable for resource-constrained deployments .

5.2 Limitations and Constraints.

The results of this work are promising but there are some important limitations that affect the generalizability and applicability of our results. The platform wants to be used for health or environment-related apps but the company does not get much demand for the same. Many countries have disasters and communicate primarily on social media in non-English languages, and no work has examined model performance on non-English text, which may be substantially different from English. The second thing is, this dataset covers past disasters because time duration has been incorporated in the data so, it may not be valid for disasters to come. Words used for disaster changes over time while social media introduces new updates and user trend behaviour. Also, new types of disasters or never-before-seen disaster features may use word patterns not present in the training data. When faced with real disasters, models relying on historical data always underfit the situation. The dataset has geographic bias. Because there are many disasters from the developed English-speaking countries. Twitter use varies a lot by region, while disasters in various geographic areas can differ. The words that we use to refer to earthquakes vary by region; then there are the terms for wildfire versus monsoon versus typhoon versus hurricane. There are other terminology for disasters too. The vocabulary that people use to describe disasters varies a lot depending on which part of the world they come from and which language they speak. We do not yet know how our model performs on disasters from different regions. This paper focuses on the binary classification of tweets into disaster and non-disaster categories. It does not classify tweets further into types of disaster, urgency and geographic impact. Benefits of fine-grained classification for Emergency Response Systems to prioritize and allocate resources based on type and severity of disaster. While the current paper does not state any approach to multi-task or hierarchical classification; it lays a foundation for such approaches. Fifth, inference of the transformer model requires the GPU acceleration to get the best performance. However, the required GPU acceleration may not be available in resource-limited emergency management situations. This will particularly be true for developing countries which are usually the most vulnerable to disasters. With DistilBERT requiring less resources, it addresses the concern. Nevertheless, deployment in resource-exhausted scenarios could use stronger reduction [5].

5.3 Ethical considerations and responsible deployment

When ML systems are to be deployed into emergency management, this raises ethical issues that go beyond performance measures. A major concern is the bias and fairness of the classifications. For instance, a classification model can have different levels of performance for different demographic

groups. Here, demographic groups can refer to tweet author characteristics like socioeconomic status, language proficiency, education level, illiteracy, etc. [4] If models consistently classify tweets from specific communities in an incorrect manner, then that community's access to emergency response degrades precisely at times when access is most useful. We recommend systematic auditing of model predictions across demographic groups (when demographic information is available); collection and evaluation of models on diverse datasets representative of many languages, dialects and geographic regions; and use of fairness constraints during training that explicitly optimize for performance equality across demographic groups, to reduce bias concerns. When classifying disasters, a compromise emerges between false negatives and false positives which creates an ethical tension. By reducing the number of false negatives, we can ensure that real disasters are not missed, and lives are saved. But minimizing false negatives generally means taking on higher rates of false positives, which uses up emergency resources in responding to a non-disaster. During the high point phase of disasters when lives are at imminent risk, false negatives are more dangerous than false positives. With a deep understanding that human operators will do the work needed to verify any alerts picked up by disaster response systems, we suggest tuning the models toward high recall with a higher false positive rate. This suggestion assumes that the organization will accept that emergency people will respond to even more false alarms and that the costs of false alarms are an acceptable cost to pay for missing disasters. Making things clear can lead to issues. Deep learning systems are often viewed as "black boxes" that make predictions that emergency responders can't understand or check. If responders get a model's classification along with the reason for that classification, they will more useful than both of over-trust and distrust. Several mitigation strategies exist. Visualization methods can show which input token most influenced a specific prediction. Having transparent base models allows for some sanity checks and alternate predictions. When predictions include confidence estimates, responders can determine the model's confidence with certain predictions. We have to create various processes where people can check model predictions and these processes will help in preserving agency to a certain extent.

Privacy issues arise due to the processing of personal tweets that may contain sensitive information about people and communities. When a disaster strikes, social media activity can reveal one's vulnerability, health status, family location information, and much more. To protect privacy we recommend that (i) data be strictly anonymised in research outputs and in deployed systems, (ii) all relevant data protection legislation be complied with, including GDPR and CCPA, (iii) only limited fields should be collected, and (iv) retention policies implemented should delete historical data after appropriate intervals.

5.4 Future Work and Research Directions

This work could be enhanced and extended by a number of important avenues for future research. Multi-task learning is a promising direction that goes beyond binary disaster classification towards predicting disaster type, urgency level, and affected geographic location simultaneously. Models like these can give emergency responders "greater situational awareness" from a single model. It reduces the need for cascading classification. By harnessing the powers of multiple architectures, we may achieve maximal accuracy and generalisation. Combining BERT, DistilBERT, and RoBERTa via voting or stacking methods would enable us to achieve such a result. In operational deployments, it could be very valuable where robustness and reliability are important. The model flags its predictions for a human expert to review and annotate. This allows the model to be improved at a low annotation cost, which is particularly useful for new disasters or languages. The integration of real-time deployment will generate comprehensive disaster response systems. There are models which processes streaming social media and feed structured classifications. To work on the integration of the incident management system, we need to address some issues. If we trained multilingual models on disaster tweets in languages other than English, their usefulness for global disaster response would increase many-fold. Fine-tuning these multilingual transformers mBERT or XLM-RoBERTa provides a starting point [3]. It would be helpful to see how language changes during a disaster to retrain the model to be more robust over time. One of the more ambitious possibilities would be to create disaster response systems that combine social media data with other information sources, such as official emergency broadcasts, sensor networks, news media, and satellite imagery.

6. Conclusion.

Disaster tweet classification accuracy and robustness can be improved using transformer-based models according to the findings of this paper. We have achieved 85.10 percent accuracy and 0.8183 F1-score using BERT through systematic experimentation, signifying an improvement of 6.5 percent F1-score over a well-tuned baseline of TF-IDF with Logistic Regression.

As a result of this method, the BERT model achieves not merely better statistics, but rather tangible operational gains. More specifically, there are 51 more disaster tweets in our validation set that BERT correctly identifies compared to the baseline. These are disaster events that will get detected through the automated system, but will otherwise get missed. After hyperparameter tuning, we found configurations that optimize for either precision or recall depending on the deployment strategy. DistilBERT gives a competitive performance [5] with a 40 percent fewer parameters [5]. It also offers 60 percent speedup. [5]. Thus, it is useful in resource-constrained deployments [5]. The confusion matrix analysis indicates that the transformer exhibits a different kind of error pattern from classical methods and that the transformer models improve performance through an entirely different error pattern. The transformer has a superior understanding of context and informal social media language than classical methods. While looking through the errors, we find that there are still some areas of difficulty that remain. They are related to the use of metaphorical language, speculative versus confirmed disaster language, and indirect references. These suggestions could arise from improving on these areas in the future. We highlighted ethical issues like bias and fairness across demographic groups, the precision-recall trade-off for emergency response, transparency and explainability for human decision-makers, and effective privacy protection for affected communities. If you adhere to these ethical issues carefully and also systematically paid attention to model transparency, and human oversight, then transformer based systems can meaningfully augment emergency response capabilities [1] and speed response to disasters, which means saving lives. Future work should focus on ensemble methods of several transformer architectures, multilingual models for global deployment, and integration with current emergency management systems for end-to-end disaster response workflows. The work done in this research creates a strong foundation for the applicability of a transformer [3] for emergency response classification and could set a good baseline for future works in this critical field.

References

[1] Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What twitter may contribute to situational awareness. *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, 1079–1088.

[2] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

[4] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 1-54.

[5] Sanh, V., Debut, L., Malmaud, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.