Authenticity-Driven Classification of Disaster-Related Tweets

Xiaoxuan Qin*
Lulin Yang*
Wendi Li*
xiq33@pitt.edu
luy30@pitt.edu
wel242@pitt.edu
University of Pittsburgh
Pittsburgh, PA, USA

Abstract

Rapid identification of disaster-related posts on social media is essential for timely situational awareness and resource allocation during emergencies. However, the brevity and informality of tweets—replete with abbreviations, misspellings, hashtags, and emojis—pose significant challenges for traditional text classifiers. We address this by framing disaster detection as a binary classification task and introducing a comprehensive evaluation of both sparse and dense text representations—ranging from bag-of-words and TF-IDF (with and without PCA) to pre-trained GloVe and BERT embeddings—paired with logistic regression, support vector machines, and feed-forward neural networks. On 7 613 labeled tweets from the Kaggle "Real or Not? NLP with Disaster Tweets" challenge, our best BERT-based model achieves an AUC of 0.87 and an accuracy of 0.82, outperforming sparse baselines by over 7% in AUC. We further conduct an error analysis to uncover common pitfalls—such as ambiguous terms like "fatal"—and demonstrate that unsupervised clustering can effectively separate natural from human-caused events among detected disaster tweets. Our findings underscore the critical role of context-aware embeddings for robust disaster monitoring and offer a practical blueprint for deploying automated tweet-classification systems in real-world crisis response.

Keywords

Disaster Detection, Tweet, Classification, Text Representation

ACM Reference Format:
Xiaoxuan Qin, Lulin Yang, and Wendi Li. 2025. Authenticity-Driven Classification of Disaster-Related Tweets. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (INFSCI 2160 Data Mining).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The rapid proliferation of social media has transformed platforms like Twitter into critical real-time channels for disseminating information during emergencies, from hurricanes and wildfires to industrial accidents and terrorist attacks. Emergency responders and policy makers increasingly rely on these decentralized data streams to gain situational awareness and allocate scarce resources effectively. However, tweets present unique challenges for automated analysis: they are exceedingly brief, often ungrammatical, and rife with slang, abbreviations, hashtags, emojis, and misspellings.

Traditional sparse text representations—Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF)—translate each tweet into a very high-dimensional vector, most of whose entries are zero. This sparsity not only inflates computational and memory costs but also lets noise from infrequent or irrelevant terms drown out critical signals. To address this, we apply Principal Component Analysis (PCA) to both BoW and TF-IDF matrices, compressing them into a lower-dimensional subspace that preserves the most informative variance. In contrast, dense text representations—whether 200-dimensional GloVe word averages or 384-dimensional BERT sentence embeddings—directly encode semantic and contextual relationships in compact vectors, avoiding manual dimensionality reduction altogether. Moreover, it remains unclear which combinations of representation and classifier yield the best performance. By systematically comparing sparse+PCA versus dense embeddings across multiple classifiers(Logistic Regression, SVM, Feed-Forward Neutral Network), we seek to determine which approach best balances interpretability, efficiency, and predictive performance in detecting real disaster tweets, what types of tweets are most prone to misclassification, and whether unsupervised clustering can meaningfully subdivide detected disaster tweets into natural versus human-caused events.

In this work, we address these questions by framing disaster detection as a binary classification task and evaluating a comprehensive suite of feature representations and models on 7,613 labeled tweets from the Kaggle "Real or Not? NLP with Disaster Tweets" dataset. Our contributions are threefold:

(1) **Empirical Benchmarking:** We provide the first head-to-head comparison of PCA-reduced BoW, PCA-reduced TF-IDF, GloVe, and BERT embeddings across multiple classifiers.
(2) **Error Analysis:** We identify common sources of misclassification—such as semantically ambiguous terms—and quantify their impact.

INFSCI 2160 Data Mining, Pittsburgh, PA

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

(3) **Clustering Extension:** We demonstrate that hierarchical and k-means clustering can distinguish between natural and human-caused disaster tweets, offering deeper insights for downstream triage.

Crucially, our results reveal that no single model dominates across all operational goals—each model-representation pair excels under different metrics (AUC, accuracy, recall, efficiency). This insight underscores the necessity of selecting models that align with specific deployment requirements in real-world crisis response. Finally, we demonstrate that dense, context-aware embeddings (especially BERT) significantly outperform sparse features in most scenarios, achieving up to 0.87 AUC and 0.82 accuracy.

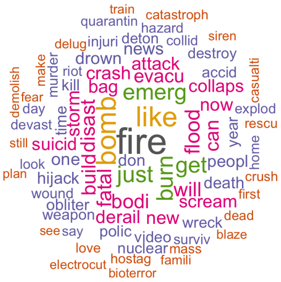

Figure 1: Overall word cloud in disaster-related Tweets

2 Related Work

The detection of disaster-related tweets has become increasingly critical as social media platforms like Twitter are now indispensable sources of real-time information during emergencies. Unlike traditional news outlets, Twitter provides decentralized, immediate, and diverse updates from affected individuals, enabling faster situational awareness for emergency responders and policymakers. Many studies have demonstrated that social media data, when effectively leveraged, can significantly enhance the speed and accuracy of crisis response. Early studies, such as Vieweg et al., analyzed Twitter communications during two natural hazard event. They showed that tweets from users in disaster areas carry rich geographic and situational information, supporting situational awareness and laying a foundation for information extraction systems [10]. Another study from Imran et al. reviewed approaches to processing social media messages during mass emergencies. It acknowledged the advantages of social media's immediacy and broad reach, while also raising concerns about data accuracy and related issues [4]. Both studies affirm the value of social media information, providing a strong foundation for this and broader research, but focus mainly on macro-level analysis with limited discussion of technical methods.

In the context of disaster detection via social media posts, one of the most crucial steps is the accurate and efficient extraction of information from raw text—that is, the choice of an appropriate text representation method. In traditional text mining, the most common approaches are Bag-of-Words (BoW) [3] and Term Frequency–Inverse Document Frequency (TF-IDF) [9], which represent text by quantifying word occurrence frequencies. While these methods are simple and computationally efficient, they create sparse, high-dimensional features and miss semantic links. To address these shortcomings, Mikolov et al.[7] introduced the concept of word embeddings, pioneering a shift towards dense, low-dimensional vector representations that capture semantic similarity. Many subsequent studies have compared different text representation methods. For instance, Deb et al. [1] compared the traditional context-free word embedding and Bert embedding, finding that contextual embeddings consistently outperformed shallow representations. Dharrao et al. [2] also found that CNN models with BERT embeddings and RMSprop optimization significantly outperformed than other methods. In general, there is broad consensus on the importance of capturing semantic information and on the strength of unsupervised text representation methods, particularly with BERT [6] [12] [8]. However, there are scholars who take a different view. Wang et al. [11] conducted a comprehensive empirical evaluation of 14 text representation methods for noisy Twitter data, finding that advances in NLP models do not necessarily lead to better performance in representing social media text. They caution that more advanced models do not always guarantee better results; text representation should be adapted to the characteristics of the task and data. Therefore, we selected both basic and advanced text representations in this study.

The selection of classification models is highly diverse, ranging from basic traditional methods such as logistic regression to deep learning approaches like Transformer-based architectures. A comparative study by Kumar et al. [5] provides a useful reference: they evaluated seven traditional machine learning classifiers and five deep learning-based models on tweet classification tasks. Their results showed that deep learning approaches generally outperform traditional methods, even with limited data. They also observed that the effectiveness of word embedding methods varies depending on the choice of classifier and application context. However, their study did not systematically examine the performance of basic text representation methods across a broader set of machine learning techniques.

Building on the insights from previous studies, it is evident that both text representation and classification model selection have been extensively investigated in the context of disaster detection via social media. While many studies have focused on comparing the performance of different methods and models, we observe a common limitation: most existing work tends to emphasize either text representation methods or classification models, but rarely both comprehensively. For instance, Deb et al. [1] compared context-free and contextualized text representations but provided limited evaluation across different classification models. Conversely, Wang et al. [11] explored a broad range of classification models but did not incorporate sentence-level BERT embeddings for text representation. From our perspective, both text representation and classification must be jointly considered to fully understand their impact on disaster detection performance. Therefore, in this study, we systematically evaluate all pairwise combinations of selected text representation methods and classification models.

3 Dataset

3.1 Data Sources and Structure

The dataset used in this study was obtained from the Kaggle competition "Natural Language Processing with Disaster Tweets," which challenges participants to predict whether a given tweet is about a real disaster event. The dataset consists of approximately 10,000 tweets, each manually annotated as either disaster-related or not, with a roughly balanced class distribution (≈ 43% positive). Each data entry includes the following variables:

- **id**: A unique identifier for each tweet.
- **keyword**: A specific keyword from the tweet (may be blank), providing additional semantic information.
- **text**: The full text of the tweet, often short, informal, and noisy.
- **target**: A binary label indicating whether the tweet refers to a real disaster (1) or not (0).

3.2 Exploratory Data Analysis

To gain preliminary insights into the dataset, we conducted an exploratory data analysis focusing on the distribution of words in the tweet texts. Figure 1 presents a word cloud generated from all tweets in the dataset. The word cloud was created using words with a minimum frequency of 20, limiting the maximum number of displayed words to 100. The words are arranged randomly. The overall word cloud shows that many common disaster-related words such as "fire," "emergency," "evacuate," and "flood" appear prominently. However, non-disaster-related words like "like" and "just" are also frequent, making it difficult to distinguish tweets about real disasters from those that are not based solely on the word cloud visualization.

To explore the potential differences more precisely, we analyzed the top 20 most frequent words separately for tweets labeled as real disasters (target = 1) and non-disasters (target = 0), as shown in Figure 2. For tweets labeled as real disasters, terms such as "fire," "news," "disaster," "California," and "emergency" dominate, aligning closely with expected disaster-related vocabulary. In contrast, for non-disaster tweets, the most frequent words include "like," "just," "new," and "love," which are more general-purpose or conversational in nature. This contrast suggests that while individual word frequencies offer stronger signals than an overall aggregation, there remains some overlap, and relying solely on basic frequency counts may still pose challenges for accurate classification.

4 Method

The overall workflow of this study is shown in Figure 3. We systematically explore combinations of text representation methods and classification models, covering a spectrum from basic to advanced techniques for both stages. Furthermore, we perform pairwise evaluations across all combinations, followed by error analysis on misclassified messages and an extension to clustering-based disaster type categorization.

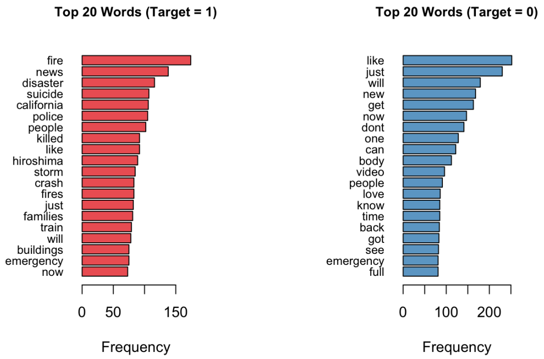

Figure 2: Top 20 most frequent words for disaster and non-disaster Tweets

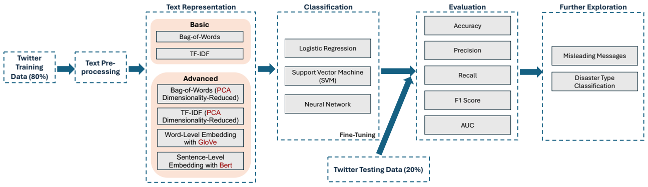
Twitter Training Data (80%)
Text Pre-processing
Text Representation
Basic
Bag-of-Words
TF-IDF
Advanced
Bag-of-Words (PCA Dimensionality-Reduced)
TF-IDF (PCA Dimensionality-Reduced)
Word-Level Embedding with GloVe
Sentence-Level Embedding with Bert
Classification
Logistic Regression
Support Vector Machine (SVM)
Neural Network
Fine-Tuning
Twitter Testing Data (20%)
Evaluation
Accuracy
Precision
Recall
F1 Score
AUC
Further Exploration
Misleading Messages
Disaster Type Classification

Figure 3: Framework for disaster Tweet classification and analysis

4.1 Pre-processing

We first combined the “keyword" and “text" columns in the dataset to create a unified text field for downstream processing. To prepare the text for feature extraction, we applied different cleaning procedures depending on the intended text representation. For traditional representations such as Bag-of-Words, TF-IDF, and GloVe embeddings, we performed extensive cleaning: removing URLs, punctuation, numbers, stopwords, emojis, and applying lowercasing, stemming, and whitespace stripping. In contrast, for BERT embeddings, we adopted a minimal cleaning approach—retaining punctuation, numbers, stopwords, and verb tenses. Since BERT, as a pre-trained contextual language model, is designed to handle syntactic and semantic nuances directly. Over-cleaning could remove valuable contextual cues and negatively impact performance.

We also examined the class distribution after pre-processing. The number of positive samples (‘target = 1‘) and negative samples (‘target = 0‘) remained relatively balanced, mitigating the risk of model bias caused by severe class imbalance.

4.2 Text Representation Method

To comprehensively evaluate model performance, we experimented with both basic and advanced text representation methods. For the basic representations, we adopted the most straightforward approaches—Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). In this setting, no additional processing such as dimensionality reduction or embedding transformation was applied. Each tweet was converted into a high-dimensional sparse feature vector based on either word counts or weighted word frequencies, providing a simple but effective baseline for subsequent comparisons.

To address the limitations of basic sparse representations and to explore richer semantic structures, we further developed a set of advanced text representations. These methods focused on two directions: dimensionality reduction and semantic embeddings. For dimensionality reduction, we applied Principal Component Analysis (PCA) to the BoW and TF-IDF matrices. Based on inspection of the principal component variance plot (scree plot), we selected the top 50 principal components that captured the majority of the information. In addition to dimensionality reduction, we implemented semantic embedding techniques at both the word and sentence levels. For word-level embeddings, each word was mapped to a 200-dimensional GloVe vector pre-trained on Twitter data. For each tweet, we computed the average of its constituent word vectors to obtain a single fixed-length vector representation. For sentence-level embeddings, each tweet was encoded as a 384-dimensional vector using the pre-trained `all-MiniLM-L6-v2` model.

4.3 Classification Model and Evaluation

We benchmarked three classifiers—Logistic Regression, Support Vector Machines (SVM), and Feed-Forward Neural Networks (NN)—on an 80/20 train–test split. Each model was trained and evaluated on all feature sets described in Section 4.2 (raw BoW, raw TF-IDF, PCA-reduced BoW, PCA-reduced TF-IDF, GloVe embeddings, and BERT embeddings) to ensure a uniform comparison.

Hyperparameters were tuned via cross-validation on the 80% training partition, and final performance was measured on the held-out 20% test set using accuracy, precision, recall, F1, and ROC-AUC.

SVM. For sparse BoW and TF-IDF features, we used a linear kernel SVM for efficiency and interpretability. For dense GloVe and BERT embeddings, we switched to a radial-basis-function kernel to capture potential nonlinear patterns. Preliminary grid searches indicated low sensitivity to the regularization constant, so we fixed $C = 1$ to avoid unnecessary tuning overhead.

Neural Networks. The conventional rule of thumb (10–50 samples per parameter) would overly restrict network size given our dataset. To preserve model capacity while mitigating overfitting, we relaxed this heuristic, applied L2 weight decay ($\lambda = 0.5$), and selected architecture via 10-fold cross-validation on the training set.

All fitted models—including the neural networks—were serialized for downstream ensemble experiments and error analysis.

4.4 Error Analysis and Clustering Extension

In the error analysis, we aim to identify which types of words most commonly mislead the classification process. To facilitate this investigation, we selected the model with high recall but low precision—indicating a high number of false positives—for further analysis. By examining the most frequent words within the false positive samples, we derived insights into common patterns that contribute to misclassification.

In clustering extension, we go beyond binary "emergency vs. non-emergency" detection and apply two complementary unsupervised methods—hierarchical clustering (average linkage) and k-means—to the subset of tweets flagged as emergencies, in order to discover interpretable incident subtypes. For tweets predicted as emergencies (SVM posterior > 0.5), we represent each message with its 200-dimensional GloVe embedding and apply two complementary unsupervised methods—average-linkage hierarchical clustering and k-means—each with k=2. Prior to clustering, embeddings are projected to two principal components for visual inspection of separation. The resulting cluster assignments are retained for downstream interpretation and validation.

Method	Model	Accuracy	Precision	Recall	F1	AUC
Bag of Words	Logistic	0.7221	0.7230	0.8392	0.7768	0.7664
	SVM	0.7168	0.7169	0.8404	0.7738	0.7634
	Neural Net	0.7727	0.7600	0.6775	0.7164	0.8128
TF-IDF	Logistic	0.7129	0.7075	0.8552	0.7744	0.7780
	SVM	0.7168	0.7169	0.8404	0.7738	0.7634
	Neural Net	0.7615	0.7491	0.6574	0.7002	0.8117
GloVe (Word-Embed)	Logistic	0.7096	0.7268	0.7948	0.7593	0.7674
	SVM	0.6978	0.6762	0.9122	0.7767	0.7645
	Neural Net	0.7254	0.6858	0.6496	0.6672	0.7812
BERT (Sent-Embed)	Logistic	0.8147	0.8317	0.8506	0.8410	0.8647
	SVM	0.7930	0.8179	0.8244	0.8211	0.8643
	Neural Net	0.8154	0.8013	0.7504	0.7750	0.8695

Table 1: Performance of disaster-related Tweets detection comparison across text representations and classifiers.

5 Evaluation Results

5.1 Empirical Benchmarking

We systematically evaluated twelve model combinations, pairing four text representation methods (Bag-of-Words, TF-IDF, Word Embedding with GloVe, and Sentence Embedding with BERT) with three classification models (Logistic Regression, Support Vector Machine, and Neural Network). For Bag-of-Words and TF-IDF, we report results based on their PCA-reduced versions, as dimensionality reduction was found to improve performance by mitigating sparsity.

Table 1 provides an overview of model performance in terms of accuracy, precision, recall, F1 score, and AUC. Across all combinations of text representations and classifiers, the models achieve consistently high recall scores, indicating that the majority of actual disaster-related tweets are successfully identified. This strong recall is particularly crucial for disaster detection, where failing to detect a true emergency tweet could result in delayed or inadequate response efforts. Even when precision varies, maintaining high recall demonstrates the system's effectiveness in prioritizing sensitivity over false negatives—an essential requirement for real-world crisis monitoring.

Within this overall trend, models leveraging dense, context-aware embeddings—particularly BERT—consistently achieve superior results across all evaluation metrics. The best-performing model, a neural network with BERT sentence embeddings, attains an AUC of 0.87 and an accuracy of 0.82, further highlighting the importance of capturing semantic context to enhance both recall and overall classification performance in disaster tweet detection.

Building on these observations, Figure 4 isolates the contribution of text representation methods, averaged across classifiers. Dense embeddings such as BERT and GloVe substantially outperform sparse representations like BoW and TF-IDF, even after dimensionality reduction (PCA). Among all approaches, BERT embeddings deliver the most consistent and significant gains across accuracy, recall, and AUC, reinforcing the importance of preserving semantic context when analyzing noisy, informal disaster-related tweets. These results further explain why models using BERT achieved superior recall and overall performance in Figure 4, highlighting text representation as a key factor in effective disaster tweet detection.

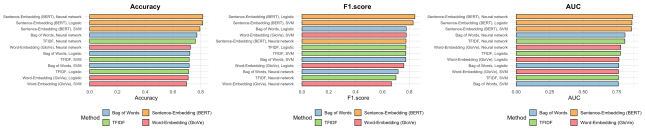

Figure 4: Comparison among text representation methods

Figure 5 further disentangles the effect of classification models, averaged across text representations. Neural networks outperform logistic regression and support vector machines (SVMs) overall, but the performance advantage is especially pronounced when paired with dense embeddings like BERT and GloVe. In contrast, when using basic sparse features, the choice of classifier has limited impact, suggesting that model complexity alone cannot compensate for poor text representations. This pattern underscores that, for disaster tweet detection, selecting a powerful text encoding method is more critical than simply choosing a more sophisticated classifier.

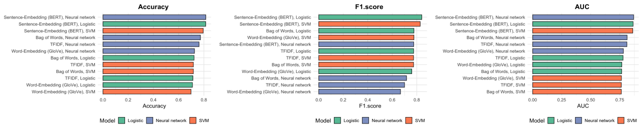

Figure 5: Comparison among classification models

5.2 Error Analysis of False Positives

The GloVe-SVM configuration was selected for error analysis because its high recall (0.91) coupled with comparatively low precision (0.67) demonstrates a pronounced tendency to generate false positives. As shown in Figure 6, an inspection of the 20 most frequent false-positive tokens revealed that 15 clearly evoke disasters (e.g., bomb, nuclear, collapse), whereas 5 are intuitively unrelated to hazardous events (e.g., sinkhole, fatal, life). Their presence highlights a noisy co-occurrence effect: during training, non-disaster terms that repeatedly appear alongside disaster vocabulary become spuriously associated with the positive class, inflating the false-positive rate.

The token fatal illustrates the problem. Although the word connotes severity, in everyday usage it often modifies abstract or metaphorical concepts—"a fatal error in the code," "fatal attraction," "fatal blow to the proposal"—that have no connection to an actual incident. Because the model relies on bag-level word counts, it lacks the contextual understanding needed to distinguish literal from figurative usage. These findings underscore the importance of incorporating context-aware features (e.g., phrase-level representations or transformer-based embeddings) or post-processing rules to mitigate spurious activations triggered by polysemous or metaphorical terms.

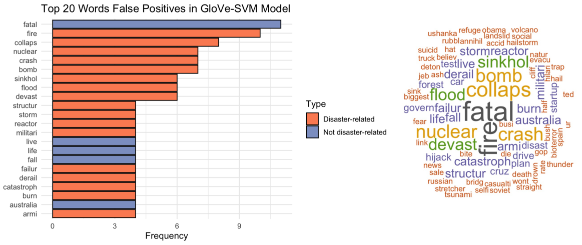

Figure 6: Top 20 most frequent words in false positive samples

5.3 Unsupervised Clustering of Disaster Sub-Types

To further explore the structure of disaster-related tweets, we applied unsupervised clustering to the subset of messages classified as emergencies. Figure 7 shows the clustering results using k-means with k=2 on the GloVe embeddings projected onto two principal components for visualization.

The left panel ("Ground Truth") plots the original emergency vs. non-emergency labels, illustrating substantial overlap and highlighting the inherent difficulty of the task. In contrast, the middle panel ("Non-Emergency vs. Emergency Clustering") shows the k-means clustering outcome: non-emergency tweets (red) are clearly separated from two distinct types of emergency tweets (green for Type 1, blue for Type 2), suggesting that the unsupervised method captures meaningful internal structure among disaster-related posts.

The right panels display the top 25 most frequent words within each cluster. Type 1 (green) is dominated by terms such as "bomb," "suicid," "fatal," and "crash," indicative of human-caused disasters like accidents, attacks, and industrial failures. Type 2 (blue) emphasizes words like "fire," "forest," "wildfire," and "burn," reflecting natural disasters such as wildfires and related environmental events.

In addition to k-means, we also experimented with hierarchical clustering, which yielded qualitatively similar groupings of natural versus human-caused disasters. Given the comparable patterns, we primarily report the k-means results for clarity and consistency.

Together, these findings demonstrate that unsupervised clustering, even without explicit disaster-type labels, can meaningfully differentiate between natural and human-caused emergencies. This suggests that embeddings capture latent thematic structures, offering a promising avenue for downstream disaster subtype classification and response prioritization.

6 Discussion

Our empirical results demonstrate several clear advantages of the proposed disaster-tweet classification pipeline. By systematically benchmarking twelve model–representation combinations, we showed that dense, context-aware embeddings—especially BERT sentence vectors—consistently yield higher discrimination (AUC = 0.87) and accuracy (= 0.82) than sparse baselines. More importantly, all of our top models maintain very high recall (> 0.90), ensuring that

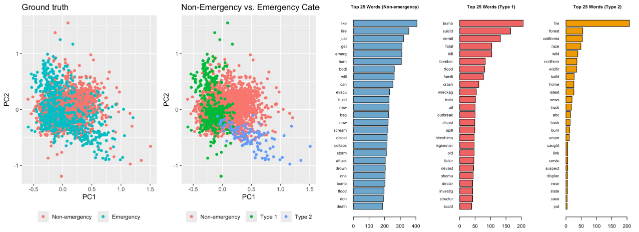

Figure 7: Separation of Natural and Human-Related Disasters via Clustering

true emergency tweets are rarely missed—a critical property for real-time crisis monitoring.

In analysing the individual components of the pipeline we gained several additional insights into the interplay between text-representation choices and classification algorithms. First, support-vector machines were the weakest performers. A plausible explanation is: SVMs are most effective in extremely high-dimensional, sparse spaces where their maximum-margin hyperplane can carve out clear class boundaries; after PCA compression or replacement with dense word- and sentence-level embeddings in this study, the feature space becomes low-dimensional and continuous, giving the algorithm far less geometric margin to exploit. Second, the GloVe representations under-performed than our expectation—even trailing the raw BoW and TF-IDF baselines. A likely explanation is twofold: (i) averaging word vectors removes word order and local context, and (ii) the 200-dimensional Twitter GloVe model was trained on generic social-media chatter rather than disaster-specific language. Together, these factors diminish its discriminative power. Third, in line with the broader literature, BERT consistently delivered the best results irrespective of the downstream classifier, reinforcing the notion that—in text mining—the choice of representation often outweighs the choice of modelling algorithm.

In generally, these findings underscore the practical importance of selecting model–representation pairs that align with specific operational goals, rather than assuming one universal "best" configuration. Therefore, we distills these findings into actionable deployment guidelines in table 2, recommending different "best" models depending on an operational goal (e.g., maximizing recall vs. minimizing inference latency).

Goal	Metric	Best Model
Maximum discrimination	AUC	BERT Neural Network
Highest predictive accuracy	Accuracy	BERT Neural Network
Balanced precision & recall	F1-score	BERT Logistic Regression
Maximum recall	Recall	GloVe SVM
Lightweight, fast inference	Efficiency	TF-IDF/BoW Logistic Regression

Table 2: Recommended models under different operational goals

Despite these strengths, several limitations warrant discussion. First, our models rely on pre-trained English embeddings and may underperform on non-English or code-mixed tweets without additional adaptation. Second, ambiguous or colloquial language (e.g., "fatal," "bomb") remains a primary source of false positives, suggesting a need for task-specific normalization or lexicon filtering. Third, while PCA reduction mitigates sparsity for BoW/TF-IDF, it also discards rare but potentially informative terms, which could impact performance on emerging or domain-specific keywords.

Looking ahead, several extensions could further improve both robustness and utility:

1. **Model fine-tuning.** Further adjust our classifiers on labeled disaster tweets to boost accuracy.
2. **Advanced and ensemble methods.** Try techniques like boosting or model stacking for more robust predictions.
3. **Misinformation filtering.** Link in external fact-check or knowledge databases to catch false or misleading posts.
4. **Time and location features.** Use tweet timestamps and geotags to improve context and detection accuracy.

In sum, our work lays a practical foundation for automated social-media–driven disaster monitoring. By choosing model–representation pairs aligned to specific operational goals (Table 2), practitioners can tailor deployments to their resource constraints and risk tolerances. Future efforts to integrate richer data sources and adaptive learning will further enhance the speed and reliability of crisis detection on social platforms.

7 Conclusion

We compared dense (GloVe, BERT) and sparse (PCA-reduced BoW/TF–IDF) text representations across three classifiers on 7,613 labeled tweets. Our key findings are: (1) BERT combined with a neural network delivers the best overall performance (AUC=0.87, accuracy=0.82); (2) all top models achieve very high recall (>0.90), crucial for ensuring emergency tweets are rarely missed; and (3) unsupervised clustering of embeddings can automatically separate natural from human-caused disasters, enabling finer-grained situational awareness without extra labels.

These results suggest that practitioners should build monitoring systems around context-aware embeddings optimized for recall, and consider embedding-based clustering to uncover latent subtypes. Future work will focus on fine-tuning models on domain data, integrating fact-checking sources, and enriching inputs with temporal and geographic metadata to further enhance robustness and timeliness.

8 Acknowledgments

1. **Xiaoxuan Qin**: Data pre-processing; classification model implementation; model fine-tuning; result interpretation; manuscript drafting.
2. **Lulin Yang**: Exploratory data analysis; feature engineering strategy; Feature engineering and regression model implementation; manuscript editing.
3. **Wendi Li**: Methodological framework design; exploratory data analysis; error analysis; data visualization; manuscript revision.
4. **Additional Thanks**: We thank the Kaggle community for providing the "Real or Not? NLP with Disaster Tweets" dataset and the authors of and the authors of `caret`, `text2vec`, and `sentence_transformers` for their open-source tooling.

References

[1] Sumona Deb and Ashis Kumar Chanda. 2022. Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data. *Machine Learning with Applications*, 7, (Mar. 2022), 100253. doi:10.1016/j.mlwa.2022.100253.

[2] Deepak Dharrao, Aadithyanarayanan Mr, Rewaa Mital, Abhinav Vengali, Madhuri Pangavhane, Satpalsing Rajput, and Anupkumar M. Bongale. 2024. An efficient method for disaster tweets classification using gradient-based optimized convolutional neural networks with BERT embeddings. *MethodsX*, 13, (Dec. 2024), 102843. doi:10.1016/j.mex.2024.102843.

[3] Zellig S. Harris. 1954. Distributional Structure. *WORD*, 10, 2-3, (Aug. 1954), 146–162. Publisher: Routledge _eprint: https://doi.org/10.1080/00437956.1954.11659520. doi:10.1080/00437956.1954.11659520.

[4] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing Social Media Messages in Mass Emergency: A Survey. *ACM Comput. Surv.*, 47, 4, (June 2015), 67:1–67:38. doi:10.1145/2771588.

[5] Abhinav Kumar, Jyoti Prakash Singh, and Sunil Saumya. 2019. A Comparative Analysis of Machine Learning Techniques for Disaster-Related Tweet Classification. In *2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129)*. ISSN: 2572-7621. (Nov. 2019), 222–227. doi:10.1109/R10-HTC47129.2019.9042443.

[6] LambaSonu, VidyarthiPranav, AggarwalMudit, GangawarPriyanshi, and MulapalliSnehita. 2025. A BERT-Based Model to Analyse Disaster's Data for Efficient Resource Management. EN. *SN Computer Science*, (Feb. 2025). Publisher: Springer Nature SingaporeSingapore. doi:10.1007/s42979-025-03720-z.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs]. (Sept. 2013). doi:10.48550/arXiv.1301.3781.

[8] A K Ningsih and A I Hadiana. 2021. Disaster Tweets Classification in Disaster Response using Bidirectional Encoder Representations from Transformer (BERT). en. *IOP Conference Series: Materials Science and Engineering*, 1115, 1, (Mar. 2021), 012032. Publisher: IOP Publishing. doi:10.1088/1757-899X/1115/1/012032.

[9] KAREN SPARCK JONES. 1972. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*, 28, 1, (Jan. 1972), 11–21. Publisher: MCB UP Ltd. doi:10.1108/eb026526.

[10] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10). Association for Computing Machinery, New York, NY, USA, (Apr. 10, 2010), 1079–1088. ISBN: 978-1-60558-929-9. doi:10.1145/1753326.1753486.

[11] Lili Wang, Chongyang Gao, Jason Wei, Weicheng Ma, Ruibo Liu, and Soroush Vosoughi. 2020. An Empirical Survey of Unsupervised Text Representation Methods on Twitter Data. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. arXiv:2012.03468 [cs], 209–214. doi:10.18653/v1/2020.wnut-1.27.

[12] Rohan Singh Wilkho, Shi Chang, and Nasir G. Gharaibeh. 2024. FF-BERT: A BERT-based ensemble for automated classification of web-based text on flash flood events. *Advanced Engineering Informatics*, 59, (Jan. 2024), 102293. doi:10.1016/j.aei.2023.102293.