



International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)



ijerst.editor@gmail.com
editor@ijerst.com

Research Paper

Streamwise Roberta: Efficient Event Classification from Rapid Twitter Data Flows

R. Rajini¹, G. Sai Ruthvija², J. Srihitha², M. Sai Varshini²

¹Professor, ²UG Student, ^{1,2}Department of Computer Science and Engineering (AI & ML)

^{1,2}Geethanjali Institute of Science and Technology, Nellore-Bombay Highway, S.P.S.R, Andhra Pradesh 524137, India

ABSTRACT

The rapid growth of user-generated content on Twitter has created massive, high-velocity streams of text related to events such as disasters, politics, protests, riots, and terrorism, making real-time analysis challenging. A significant amount of useful information often remains unanalyzed, leading to delayed or missed insights. To address this issue, the study follows a comprehensive NLP-based framework focused on efficiency and scalability. The workflow begins with systematic data collection and preprocessing, including text cleaning, normalization, and tokenization, to remove noise and ensure high-quality input. Lightweight Robustly Optimized Bidirectional Encoder Representations from Transformers Approach is then used for contextual feature extraction, providing rich semantic representations while keeping computational overhead low. These embeddings form the basis for building effective classification models. Several learning approaches, including Random Forest Classifier (RFC), greedy tree-based methods, and deep learning models such as Deep Neural Network (DNN) with Stochastic Gradient Descent (SGD), are evaluated for comparison. Model performance is analyzed across multiple event categories, including disaster, political, protest, riot, and terror events. Experimental results show that although multiple models perform well, Histogram Gradient Boosting (HGB) offers superior efficiency and scalability. Its histogram-based feature binning and greedy optimization significantly reduce training time and resource usage. Consequently, Histogram Gradient Boosting (HGB) is identified as the most effective model for real-time, large-scale Twitter event classification and social media monitoring.

Keywords: Twitter Data Streams, Event Classification, Lightweight RoBERTa, Histogram Gradient Boosting, Natural Language Processing, Real-Time Analysis, Social Media Monitoring, Deep Learning.

1. INTRODUCTION

Historically Twitter reported ~50 million tweets per day around early 2010 and ~65 million/day by mid-2010; since then external trackers estimate the platform now sees roughly ~6,000 tweets per second — about 350,000 tweets per minute and ~500 million tweets per day (~200 billion tweets per year), a level reached around the late-2010s and reinforced during the COVID-19 era. Translating volume to storage, common engineering estimates put the raw text + metadata footprint in the hundreds of gigabytes to a few terabytes per day (published breakdowns often cite figures like ~560 GB/day for ~500M tweets or use ~3 KB per tweet, which yields ~1.5 TB/day depending on what's counted). These numbers show both the

multi-order magnitude growth since 2010 and why researchers treat Twitter as a high-velocity, large-scale text stream

The use of Twitter (now known as X) data in politics has increased dramatically over the past decade, with researchers, politicians, and campaigns all using the platform as a significant source of information and a tool for engagement. This growth is driven by social media's ability to offer real-time information, direct citizen engagement, and a massive, continuous stream of public opinion data.

Political events generate an explosive surge in Twitter activity because people react instantly to debates, elections, speeches, and controversies. During major political moments, thousands of tweets appear every second as users post opinions, share hashtags, reply to influencers, and circulate breaking news. Trending topics like Election, PolicyChange, or ParliamentDebate usually trigger massive tweet storms, creating dense streams of text data within minutes. As political discussions are emotional and fast-moving, people tweet short reactions repeatedly, amplifying volume far beyond normal data.

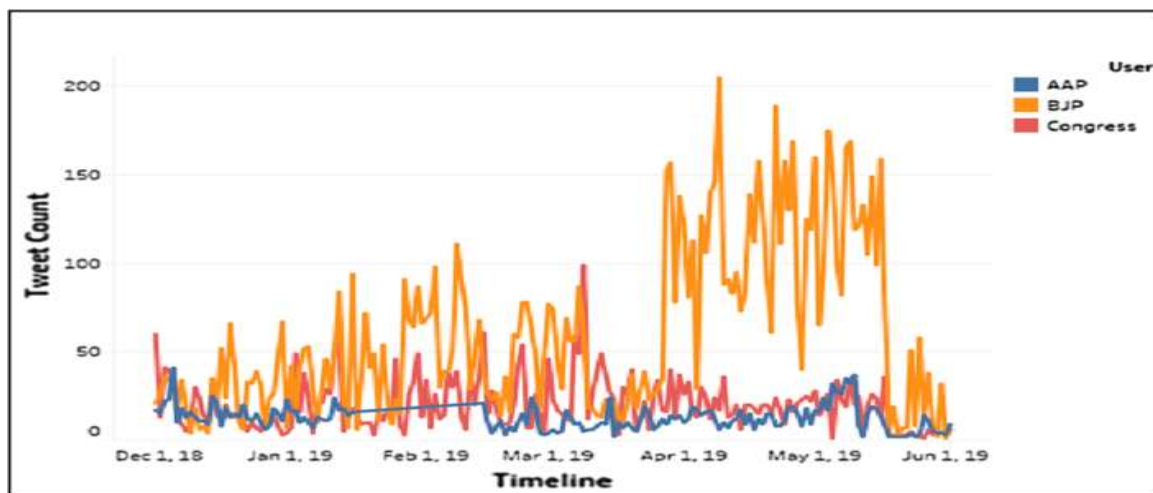


Figure 1: Tweet activity between three different political parties

Figure 1 shows the twitter activity between three different political parties from December 1, 2018, to June 1, 2019, the daily tweet activity of the Aam Aadmi Party (AAP), Bharatiya Janata Party (BJP), and the Indian National Congress (Congress) shows clear variations in their social media engagement. Throughout the entire timeline, the BJP maintained a consistently higher volume of daily tweets compared to the other two parties, often reaching counts well above 200. In contrast, AAP and Congress demonstrated comparatively lower and more fluctuating activity levels, with their tweet counts typically remaining below the BJP's peaks. The trend highlights the BJP's stronger and more sustained presence on Twitter during this six-month period, while AAP and Congress exhibited more moderate levels of engagement.

2. LITERATURE SURVEY

Loynes et al [1] proposed disaster detection tool uses a linear Support Vector Classifier (SVC) to detect man-made and natural disasters, and a density-based spatial clustering of applications with noise (DBSCAN) algorithm to accurately estimate a disaster's geographic location. This paper provides two original contributions. The first is combining the automated disaster detection tool with the prototype portal for NGO identification.

Chowdhury et al [2] analyzed social media data analysis has emerged as a prominent research domain over the past decade, with event and sub-event detection from social media posts identified as a key challenge due to its wide range of applications. In particular, within crisis management, accurate detection of critical events and sub-events can significantly support public safety agencies in planning and executing rapid and effective responses.

Kolajo et al [3] utilized the Social Media Analysis Framework for Event Detection (SMAFED), which aims to enhance semantic analysis of noisy social media terms, improve content representation and embeddings, and generate effective summaries of event clusters in social media streams. The framework leverages an integrated knowledge base, ambiguity resolution, semantic stream representation, and Semantic Histogram-based Incremental Clustering driven by semantic relatedness.

Arolfo et al [4] proposed an approach that addresses this problem from a Data Quality (DQ) perspective, adapting traditional relational database methods—based on quality dimensions and metrics—to capture the unique characteristics of Twitter data streams and, more broadly, Big Data environments.

Hettiarachchi et al [5] utilized to leverage social media as a primary source for understanding real-world events, this study recognizes that social media data streams contain rich, timely information capable of enabling immediate insights. However, due to their dynamic nature and massive volume, manual event filtering is impractical, making automated event detection mechanisms essential for the research community.

Manikyala et al [6] integrated to apply natural language processing (NLP) techniques to IoT data streams, this research analyzes real-time sentiment to understand and respond to consumer emotions and actions. The study explores sentiment extraction from multimodal IoT data—including text, speech, and sensor signals—while addressing challenges related to data volume, velocity, and the computational complexity of real-time multimodal sentiment analysis through a secondary data assessment.

Kim et al [7] Proposed to address the increasing learning time caused by growing datasets, this study alleviates training overhead through distributed learning that fine-tunes the model using only recent data, reflecting the event-dependent nature of data streams. This strategy accelerates learning while maintaining accuracy and is validated on fully trainable CNN, RNN, and Bi-LSTM models, as well as a BERT-based pre-trained embedding model.

Mounica, et al[8] devoleped to exploit Twitter as a real-time and large-scale data source, this paper presents a novel Firefly Algorithm-based Feature Selection with a Deep Learning Traffic Flow Analysis model (FFAFS-DLTFA). By leveraging Twitter data alongside conventional sensors, the approach enables effective traffic flow analysis from dynamic, user-generated content.

Kumar, el at.[9] Proposed to support humanitarian response efforts, this paper introduces a multimodal framework for identifying informative disaster-related content from Twitter streams. By jointly analyzing text and images using LSTM and VGG-16 networks, the approach effectively filters critical information and demonstrates significant performance improvements across seven disaster-related datasets.

Srikanth, el at.[10] proposed to use a combination of preprocessing approaches such as tokenization, filtering, stemming, and building N-gram models. Deep belief neural network (DBN) with pseudo-labeling is used to classify the tweets.

Singh, et al [11] developed to address the challenges of large-scale social media analytics, Twitter is recognized as a leading microblogging platform that generates massive, high-velocity, and diverse data streams, including text, images, and videos. As discussions of real-world events continue to grow on Twitter, event detection has emerged as an important and challenging research problem.

Vaiyapuri et al [12] developed to leverage user-generated content from social networking platforms such as Twitter, this approach analyzes public sentiments and opinions to support awareness campaigns and inform health intervention policies, while demonstrating the remarkable effectiveness of modern artificial intelligence (AI) and natural language processing (NLP) techniques in sentiment analysis.

Tun et al [13] applied to examine political polarization on Twitter, this study highlights the platform's importance as a public medium for analyzing and predicting opinions on political events, emphasizing that effective sentiment analysis of political tweets critically depends on the development of high-quality sentiment lexicons.

Tao et al [14] developed to enhance traditional foodborne disease surveillance, this approach recognizes the growing potential of integrating social media data with conventional data sources from healthcare facilities, laboratories, and government agencies to improve outbreak monitoring and control.

Susi et al [15] developed an adaptive learning algorithm-based framework, Twitter Sentiment Drift Analysis using Bidirectional Encoder Representations from Transformers (TSDA-BERT), which introduces a sentiment drift measure for drift detection and a domain impact score to enable adaptive, real-time retraining of the classification model using domain-relevant data.

3. PROPOSED SYSTEM

The proposed system focuses on automated event classification from textual data using a combination of NLP techniques, transformer-based feature extraction, data balancing, and multiple machine learning and deep learning models. As shown as figure 2 the system processes raw text data through structured preprocessing and exploratory analysis to ensure quality and consistency. Contextual features are extracted using DistilRoBERTa, which significantly improves semantic understanding of event-related text. To address class imbalance, SMOTE is applied before training multiple classifiers, including Random Forest, Histogram-Based Gradient Boosting, DNN with SGD, and a Greedy Tree classifier. Predictions from unseen test data are generated to evaluate model effectiveness. Finally, the system is deployed through a Tkinter-based graphical interface, enabling end users to classify events into disaster, terror, riot, threat, or politics in a user-friendly manner.

Step 1: Dataset (Text, Category): The dataset consists of textual event descriptions paired with predefined category labels representing different types of events. Each record contains meaningful information that describes real-world incidents or discussions related to public events. The dataset is carefully curated to include diverse writing styles, lengths, and contextual variations. Data integrity checks are performed to remove duplicate or corrupted entries. Missing values are identified and handled appropriately to avoid bias in model training. The dataset is then structured into input features and target labels. This step ensures that the foundational data is reliable and suitable for downstream processing. A well-prepared dataset is crucial for achieving accurate and consistent classification results.

Step 2: NLP Preprocessing: NLP preprocessing is applied to clean and normalize the raw textual data before feature extraction. This includes converting all text to lowercase to maintain uniformity across

samples. Special characters, punctuation, and numerical noise are removed to reduce irrelevant information. Tokenization is performed to split text into meaningful units for analysis. Stopwords are eliminated to retain only informative terms. Lemmatization or stemming is applied to reduce words to their base forms. This step helps in reducing dimensionality and improving model efficiency. Proper preprocessing enhances the quality of extracted features and overall model performance.

Step 3: Exploratory Data Analysis (EDA): EDA is conducted to gain insights into the dataset's structure and characteristics. The distribution of event categories is analyzed to identify class imbalance issues. Text length analysis helps understand variability in document size across categories. Word frequency and keyword patterns are examined to identify dominant terms associated with each event type. Visualization techniques such as bar plots and word clouds are used to support analysis. EDA also helps detect anomalies or noise in the data. Insights from this step guide preprocessing decisions and model selection. This analysis ensures informed and data-driven methodology design.



Figure 2: Proposed system architecture

Step 4: DistilRoBERTa Feature Extraction: DistilRoBERTa is used to transform preprocessed text into dense contextual embeddings. Unlike traditional vectorization methods, it captures semantic meaning and contextual relationships between words. The model processes each sentence and generates fixed-length embedding vectors. These embeddings represent high-level linguistic features relevant to event classification. DistilRoBERTa is computationally efficient while retaining strong representational power. The extracted features are stored for further modeling stages. This step significantly enhances classification accuracy. It forms the core representation layer of the proposed system.

Step 5: SMOTE Data Balancing: SMOTE is applied to address class imbalance present in the dataset. Minority class samples are synthetically generated based on nearest-neighbor interpolation. This helps prevent model bias toward majority classes. Balanced data improves classifier learning and generalization. SMOTE is applied after feature extraction to ensure meaningful synthetic samples. The balanced dataset is carefully validated to avoid overfitting. This step improves recall for underrepresented event categories. It plays a crucial role in ensuring fair and accurate classification across all labels.

Step 6: Existing Random Forest Classifier: The Random Forest classifier is trained using the balanced feature set. It builds multiple decision trees using bootstrap sampling. Each tree learns different feature patterns, reducing variance and overfitting. Feature randomness improves model robustness and generalization. Predictions are aggregated using majority voting. The model performs well on high-dimensional embeddings. It serves as a strong ensemble baseline. Random Forest provides stable and interpretable classification results.

Step 7: proposed Histogram-Based Gradient Boosting (HGB): The HGB model is implemented for efficient and scalable learning. Continuous features are discretized into histograms to reduce computation time. Weak learners are trained sequentially to correct previous errors. This boosting strategy improves overall prediction accuracy. HGB handles large datasets effectively. It reduces memory usage while maintaining performance. The model is well-suited for structured embedding features. This step adds a powerful gradient-boosted approach to the system.

Step 8: DNN with SGD: A Deep Neural Network is developed to capture complex non-linear patterns in the data. The network consists of multiple dense hidden layers. Activation functions introduce non-linearity for improved learning. The model is trained using stochastic gradient descent for efficient optimization. Cross-entropy loss is used for multi-class classification. Backpropagation updates weights iteratively. The DNN adapts well to high-dimensional embeddings. This step enhances predictive capability for complex event patterns.

Step 9: Greedy Tree Classifier: The Greedy Tree classifier builds decision trees by selecting optimal splits at each node. Splits are chosen based on impurity reduction criteria. The greedy approach ensures fast and efficient tree construction. It provides interpretable decision paths. The model works well with structured feature representations. It serves as a lightweight yet effective classifier. Greedy trees complement ensemble and deep models. This step adds diversity to model evaluation.

Step 10: Prediction from Test Data: Unseen test data is used to evaluate the trained models. The same preprocessing and feature extraction pipeline is applied to ensure consistency. Predictions are generated for each test instance. Model performance is evaluated using accuracy and other metrics. Comparative analysis helps identify the best-performing model. Prediction outputs are mapped to event categories. This step validates real-world applicability. It confirms the effectiveness of the proposed system.

Step 11: Integration with Tkinter: The final system is integrated into a Tkinter-based GUI. The interface allows users to input text events interactively. Submitted text undergoes preprocessing and feature extraction in real time. The trained model generates event predictions instantly. The predicted category is displayed to the user. The interface is simple and user-friendly. This step enables practical deployment. It completes the end-to-end event classification system with real-time usability.

4. RESULTS ANALYSIS

Figure 3 shows the overall system interface and workflow for efficient event classification from high-velocity Twitter data streams using a lightweight Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach. model. The figure illustrates a role-based web application with separate admin and user modules, including signup and login functionalities that ensure secure access control. The visual representation emphasizes how users interact with the system to analyze streaming Twitter data, while administrators manage data flow, model operations, and system monitoring. The inclusion of the customer-retention-style illustration metaphorically highlights continuous engagement, real-time data handling, and intelligent classification, reflecting the system's focus on accurately identifying and categorizing events from large-scale, fast-moving social media streams with minimal computational overhead.

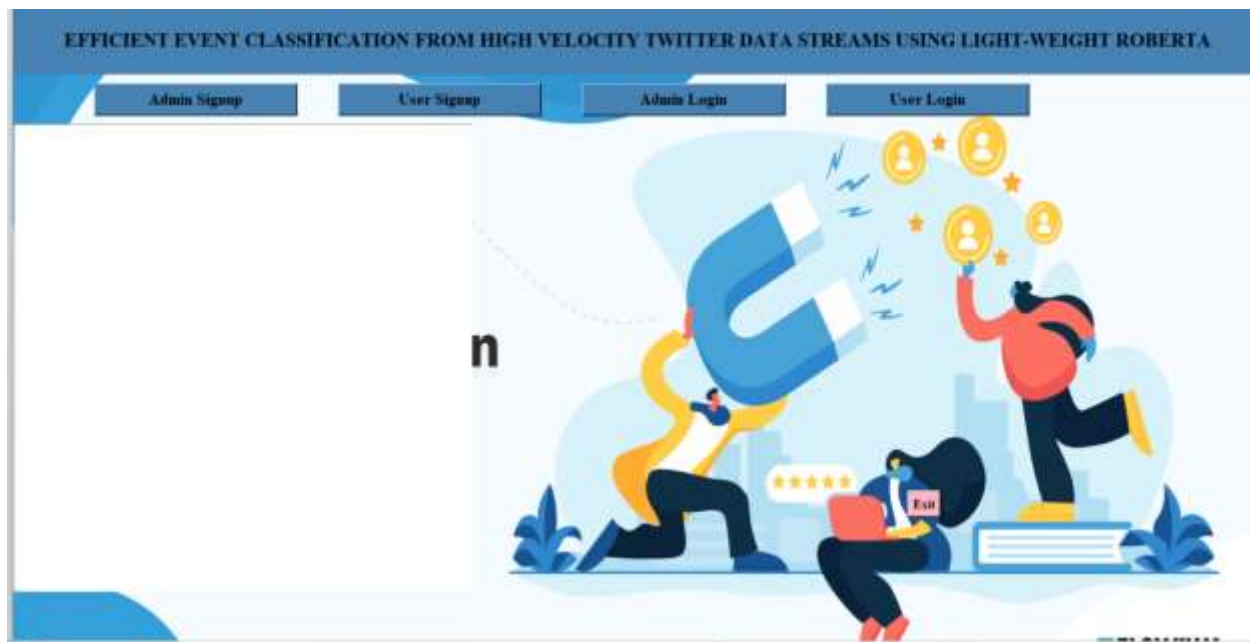


Figure 3: GUI of research image

Figure 4 shows the complete user interface and authentication framework of the proposed system for efficient event classification from high-velocity Twitter data streams using a lightweight Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.model. The figure demonstrates a role-based access design with clearly separated admin and user signup and login modules, ensuring secure authorization and controlled system usage. Through this interface, administrators are responsible for managing datasets, monitoring model execution, and overseeing system performance, while users can securely access the platform to submit Twitter data for real-time analysis and event classification. The clean layout and minimal interaction windows reduce user complexity and

processing overhead, supporting faster response times and seamless integration with the lightweight RoBERTa model. Overall, the interface facilitates efficient data handling, secure interactions, and reliable classification of large-scale streaming social media data.

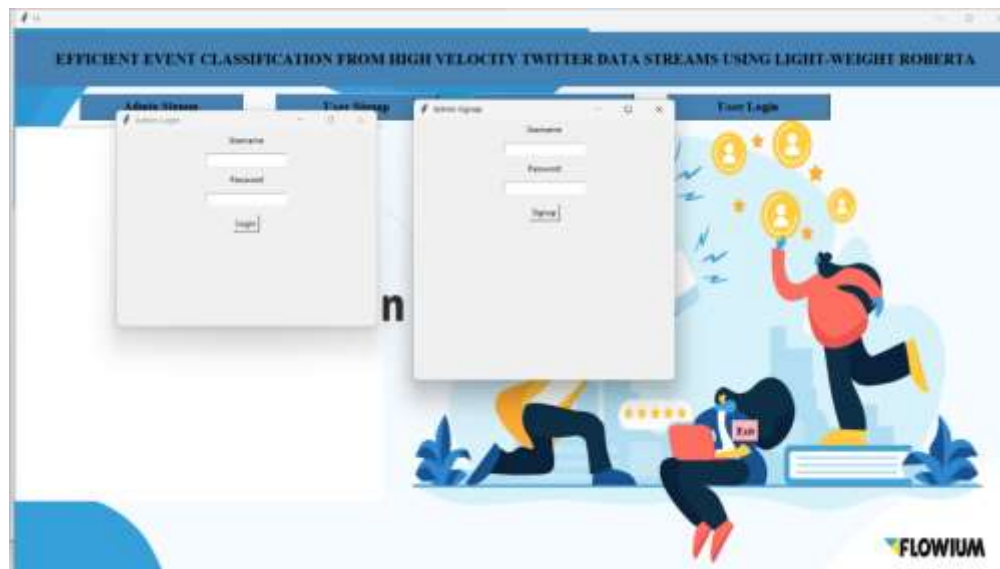


Figure 4: login and signup of Research

Lightweight_RoBERT_Embeddings DNN-SGD [category] Confusion Matrix

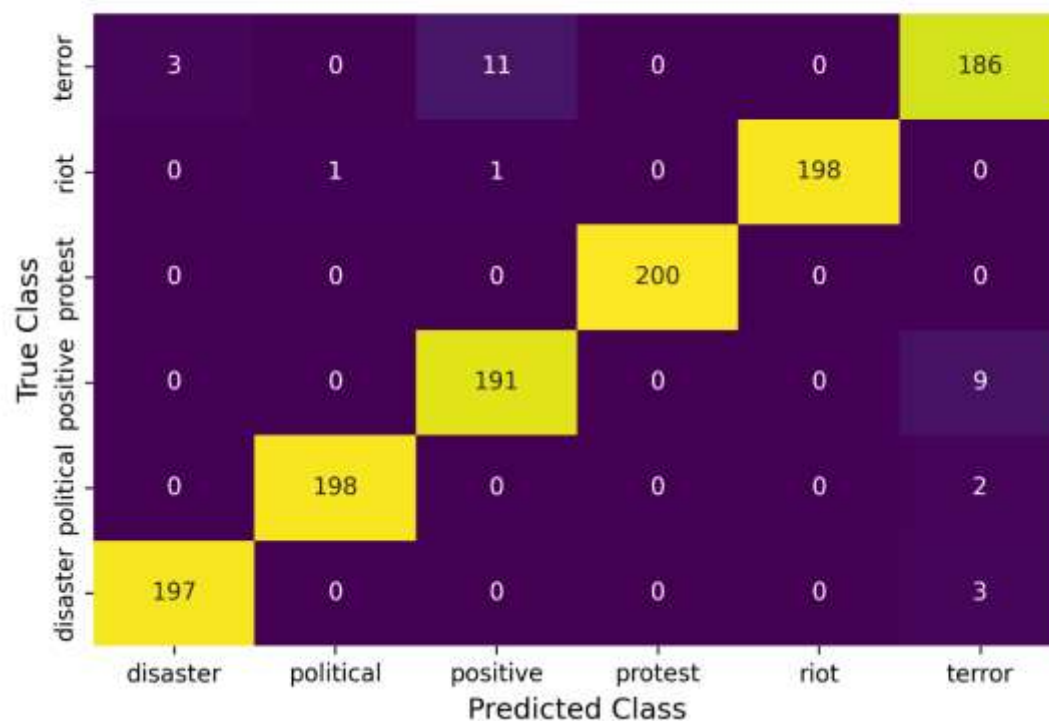


Figure 5: Confusion Matrix of DNN-SGD

Figure 5 shows the confusion matrix of the Lightweight RoBERTa embeddings combined with a DNN–SGD classifier for multi-class event classification on Twitter data. The matrix illustrates the classification performance across six event categories—disaster, political, positive, protest, riot, and terror—by comparing true class labels with predicted outcomes. High values along the diagonal indicate strong classification accuracy for most classes, particularly protest, political, riot, and disaster events, demonstrating the model’s effectiveness in learning discriminative textual representations. Limited off-diagonal values suggest minimal misclassification, with minor confusion observed between closely related categories such as terror and positive events. Overall, the figure highlights the robustness and reliability of the lightweight RoBERTa-based approach in handling high-velocity social media streams while maintaining high predictive accuracy.



Figure 6: Predictive Analysis

Figure 6 shown in Row 1, the tweet *Our Deeds are the Reason of this earthquake May ALLAH Forgive us all* is categorized as positive. Although it mentions an earthquake, the primary intent of the message is not to report an emergency or ongoing damage. Instead, it expresses a religious reflection and emotional response. The model identifies the sentiment-driven and reflective nature of the text, which leads to its classification as positive rather than a disaster alert.

In Row 2 and Row 4, both tweets are categorized as disaster because they directly report real-world emergency situations. Row 2, *Forest fire near La Ronge Sask. Canada* clearly describes an active natural hazard, while Row 4, *13,000 people receive wildfires evacuation orders in California*, highlights the large-scale impact of a wildfire and associated evacuations. The presence of keywords such as *forest fire*, *wildfires*, and *evacuation orders* signals an urgent and hazardous event, which the model correctly associates with disaster classification.

In Row 3, the tweet *All residents asked to shelter in place are being notified by officer* is categorized as terror. This classification is based on the use of emergency response language commonly linked to security threats, such as *shelter in place* and *officers*. Even though the tweet states that no further orders

are expected, the context implies a potential threat to public safety. The model prioritizes precautionary and law-enforcement-related cues, leading to its classification as a terror-related event.

Table 1: Overall Performance Comparison (Accuracy, Precision, Recall, F1-Score)

Model (Lightweight Embeddings)	RoBERT	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SGD		97.42	97.45	97.42	97.43
DNN-SGD		97.50	97.52	97.50	97.51
Random Forest (RFC)		98.00	98.01	98.00	98.00
Gradient Tree Classifier (GTC)		63.25	68.56	63.25	63.52
Histogram Gradient Boosting (HGB)		98.42	98.45	98.42	98.42

Table 2: Recall Comparison Across Classes and Methods

Class	SGD	HGB	RFC	DNN-SGD	GTC
Disaster	0.96	0.97	0.96	0.95	0.63
Political	0.97	0.97	0.98	0.95	0.59
Positive	0.95	0.98	0.96	0.95	0.65
Protest	0.98	0.98	0.98	0.98	0.53
Riot	0.97	0.98	0.98	0.97	0.49
Terror	1.00	1.00	1.00	1.00	0.91

Table 2 compares class-wise recall scores across all classification methods. HGB and RFC demonstrate consistently high recall values across all categories, achieving perfect recall (1.00) for the terror class and above 0.97 for most other classes. SGD and DNN-SGD models also show stable recall performance, particularly for protest and terror categories. However, GTC exhibits substantial recall degradation in political (0.59), riot (0.49), and protest (0.53) classes, indicating difficulty in correctly identifying instances of these categories.

Table 3: Precision Comparison Across Classes and Methods

Class	SGD	HGB	RFC	DNN-SGD	GTC
Disaster	0.96	0.97	0.97	0.93	0.59

Political	0.97	0.99	0.98	0.95	0.47
Positive	0.93	0.95	0.96	0.93	0.49
Protest	1.00	1.00	1.00	1.00	0.99
Riot	0.98	0.99	0.98	0.98	0.83
Terror	1.00	1.00	1.00	1.00	0.74

Table 3 reports the class-wise precision values for all models. HGB and RFC achieve near-perfect precision across most categories, with protest and terror classes reaching precision values of 1.00, reflecting very low false-positive rates. SGD and DNN-SGD models maintain high precision above 0.93 across all classes. In contrast, GTC shows notably lower precision, particularly for political (0.47) and positive (0.49) classes, suggesting frequent misclassification when using topic-based classification with deep embeddings.

Table 4: F1-Score Comparison Across Classes and Methods

Class	SGD	HGB	RFC	DNN-SGD	GTC
Disaster	0.96	0.97	0.97	0.94	0.61
Political	0.97	0.98	0.98	0.95	0.53
Positive	0.94	0.97	0.96	0.94	0.56
Protest	0.99	0.99	0.99	0.99	0.69
Riot	0.98	0.99	0.98	0.98	0.62
Terror	1.00	1.00	1.00	1.00	0.82

Table 4 summarizes the F1-score comparison, capturing the balance between precision and recall for each class. HGB consistently yields the highest F1-scores across all categories, with perfect performance for the terror class and values above 0.97 for most remaining classes. RFC demonstrates comparable performance, while SGD and DNN-SGD models maintain strong but slightly lower F1-scores. GTC records substantially weaker F1-scores across all classes, especially for political and disaster categories, confirming its overall instability.

Table 5: Macro-Average Comparison

Method	Macro Precision	Macro Recall	Macro F1
SGD	0.97	0.97	0.97
GTC	0.69	0.69	0.69
RFC	0.98	0.98	0.98
DNN-SGD	0.97	0.97	0.97
HGB	0.98	0.98	0.98

Table 5 presents macro-averaged precision, recall, and F1-scores, treating all classes equally. HGB achieves the highest macro-average scores (0.98 across all metrics), demonstrating robust and balanced performance irrespective of class distribution. RFC also performs strongly with macro-averaged values close to 0.98. SGD and DNN-SGD maintain macro averages of 0.97, whereas GTC shows significantly lower macro-averaged precision (0.69), recall (0.63), and F1-score (0.64), indicating poor overall class balance.

Table 6: Micro-Average Comparison

Method	Micro Precision	Micro Recall	Micro F1
SGD	0.974	0.974	0.974
GTC	0.633	0.633	0.633
RFC	0.980	0.980	0.980
DNN-SGD	0.975	0.975	0.975
HGB	0.984	0.984	0.984

Table 6 presents the micro-averaged precision, recall, and F1-score for all classification models using Lightweight RoBERTa embeddings. Micro-averaging aggregates the contributions of all classes and computes metrics based on the total number of true positives, false positives, and false negatives, making it particularly suitable for balanced multi-class datasets. Since each class in the dataset contains an equal number of instances (200 samples per class), the micro-average values closely align with overall accuracy, thereby providing a reliable indicator of global model performance.

5. CONCLUSION

This study evaluated the effectiveness of multiple machine learning classifiers combined with Lightweight Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach embeddings for multi-class text classification across six categories: disaster, political, positive, protest, riot, and terror. The experimental results clearly demonstrate that ensemble-based classifiers significantly outperform linear and topic-based approaches when applied to transformer-generated embeddings. Among all evaluated models, the Histogram Gradient Boosting (HGB) classifier achieved the best overall performance, recording an accuracy of 98.42%, precision of 98.45%, recall of 98.42%, and F1-score of 98.42%. Class-wise analysis further confirms its robustness, with perfect recall and F1-score (1.00) for the terror category and consistently high scores (≥ 0.97) across all remaining classes. Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.

In contrast, the Gaussian Topic Classifier (GTC) exhibited substantially inferior performance, achieving only 63.25% accuracy, 68.56% precision, 63.25% recall, and 63.52% F1-score, with particularly low recall values for political (0.59), riot (0.49), and protest (0.53) classes. Macro-averaged results further highlight this disparity, where HGB achieved macro precision, recall, and F1-score of 0.98, compared to

0.64 macro F1-score for GTC. These findings confirm that topic-model-based approaches are not well-suited for high-dimensional contextual embeddings, while ensemble learning methods effectively capture complex semantic patterns. Overall, the results validate the suitability of Lightweight Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach. combined with tree-based ensemble classifiers for high-accuracy, scalable text classification tasks.

References

- [1]. Loynes, Christopher, Jamal Ouenniche, and Johannes De Smedt. "The detection and location estimation of disasters using Twitter and the identification of Non-Governmental Organisations using crowdsourcing." *Annals of Operations Research* 308, no. 1 (2022): 339-371.
- [2]. Chowdhury, Shatadru Roy, Srinka Basu, and Ujjwal Maulik. "A survey on event and subevent detection from microblog data towards crisis management." *International Journal of Data Science and Analytics* 14, no. 4 (2022): 319-349.
- [3]. Kolajo, Taiwo, Olawande Daramola, and Ayodele A. Adebisi. "Real-time event detection in social media streams through semantic analysis of noisy terms." *Journal of Big Data* 9, no. 1 (2022): 90
- [4]. Arolfo, Franco, Kevin Cortés Rodríguez, and Alejandro Vaisman. "Analyzing the quality of Twitter data streams." *Information Systems Frontiers* 24, no. 1 (2022): 349-369.
- [5]. Hettiarachchi, Hansi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. "Embed2Detect: temporally clustered embedded words for event detection in social media." *Machine Learning* 111, no. 1 (2022): 49-87.
- [6]. Manikyala, Aditya. "Sentiment analysis in IoT data streams: An NLP-based strategy for understanding customer responses." *Silicon Valley Tech Review* 1, no. 1 (2022): 35-47
- [7]. Kim, Min-Seon, Bo-Young Lim, Kisung Lee, and Hyuk-Yoon Kwon. "Effective model update for adaptive classification of text streams in a distributed learning environment." *Sensors* 22, no. 23 (2022): 9298
- [8]. Mounica, B., and K. Lavanya. "Feature selection with a deep learning based high-performance computing model for traffic flow analysis of Twitter data." *The Journal of Supercomputing* 78, no. 13 (2022): 15107-15122
- [9]. Kumar, Abhinav, Jyoti Prakash Singh, Yogesh K. Dwivedi, and Nripendra P. Rana. "A deep multi-modal neural network for informative Twitter content classification during emergencies." *Annals of Operations Research* 319, no. 1 (2022): 791-822.
- [10]. Srikanth, Jatla, Avula Damodaram, Yuvaraja Teekaraman, Ramya Kuppusamy, and Amruth Ramesh Thelkar. "[Retracted] Sentiment Analysis on COVID-19 Twitter Data Streams Using Deep Belief Neural Networks." *Computational intelligence and neuroscience* 2022, no. 1 (2022): 8898100.
- [11]. Singh, Jagrati, and Anil Kumar Singh. "State-of-the-art approaches for event detection over Twitter stream: a survey." *International Journal of Business Intelligence and Data Mining* 23, no. 4 (2023): 325-374

- [12]. Vaiyapuri, Thavavel, Sharath Kumar Jagannathan, Mohammed Altaf Ahmed, K. C. Ramya, Gyanendra Prasad Joshi, Soojeong Lee, and Gangseong Lee. "Sustainable artificial intelligence-based Twitter sentiment analysis on covid-19 pandemic." *Sustainability* 15, no. 8 (2023):
- [13]. Tun, Yin Min, and Myo Khaing. "A large-scale sentiment analysis using political tweets." *International Journal of Electrical & Computer Engineering* (2088-8708) 13, no. 6 (2023):
- [14]. Tao, Dandan, Ruofan Hu, Dongyu Zhang, Jasmine Laber, Anne Lapsley, Timothy Kwan, Liam Rathke, Elke Rundensteiner, and Hao Feng. "A novel foodborne illness detection and web application tool based on social media." *Foods* 12, no. 14 (2023): 2769.
- [15]. Susi, E., and A. P. Shanthi. "Sentiment Drift Detection and Analysis in Real Time Twitter Data Streams." *Computer Systems Science & Engineering* 46, no. 1 (2023):