An Efficient SBERT-Enhanced RoBERTa Feature Extraction Model with Machine Learning Classifiers for Real-Time Political Event Detection on Twitter

Goutham Kunamalla[1], Mohammad Afroz[2], Macherla Pavithra[2], Mahankali Shivani[2], Eppalapally Sathwik[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

To Cite this Article
Goutham Kunamalla, Mohammad Afroz, Macherla Pavithra, Mahankali Shivani, Eppalapally Sathwik, "An Efficient SBERT-Enhanced RoBERTa Feature Extraction Model with Machine Learning Classifiers for Real-Time Political Event Detection on Twitter", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04, April 2026,pp: 470-479, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04.pp470-479

Submitted: 28-02-2026 Accepted: 02-04-2026 Published: 10-04-2026

ABSTRACT

The rapid evolution of Twitter as a real-time information-sharing platform has led to an unprecedented surge in short-text data reflecting ongoing political and social events. Analyzing this vast and unstructured data manually is both inefficient and unreliable, particularly for time-critical applications such as crisis response and governance. Conventional machine learning techniques often fail to capture deep semantic and contextual relationships within textual data, thereby limiting their effectiveness in complex event classification tasks. To address these limitations, this study proposes an efficient political event detection framework that integrates SBERT-enhanced Lightweight Robustly Optimized Bidirectional Encoder Representation from Transformer (RoBERTa) for feature extraction with advanced machine learning classifiers. Initially, raw tweet data undergoes comprehensive Natural Language Processing (NLP) preprocessing, including tokenization, stop-word removal, normalization, and lemmatization, ensuring high-quality input. Exploratory Data Analysis (EDA) is conducted to uncover data distribution and linguistic patterns. Subsequently, semantic-rich embeddings are generated using a lightweight SBERT-based RoBERTa model, enabling context-aware feature representation. These features are then utilized by multiple classifiers, including Stochastic Gradient Descent (SGD), Histogram Gradient Boosting (HGB), Random Forest Classifier (RFC), and Greedy Tree Classifier (GTC). Furthermore, a Deep Neural Network (DNN) is incorporated for intermediate feature learning, followed by an optimized SGD classifier to enhance predictive performance. The proposed system classifies tweets into six categories: disaster, political, positive, protest, riot, and terror. Experimental results demonstrate superior accuracy, scalability, and generalization compared to traditional methods, highlighting its effectiveness for real-time political event detection and decision-support systems.

Keywords: Political Event Detection, NLP, RoBERTa, SBERT, DNN, Machine Learning, Twitter Analysis, Text Classification, Real-Time Systems.

1. INTRODUCTION

Twitter has emerged as a significant digital communication platform where users continuously share real-time updates, opinions, and discussions related to global developments. Among these, political events constitute a major portion of online discourse, significantly shaping public perception, policy decisions, and international relations. Recent studies indicate that more than 65% of internet users depend on social media platforms such as Twitter for accessing news, with political topics consistently

ISSN: 3049-0952
www.jsetms.com

ranking among the most actively discussed and engaged categories. As shown as figure 1 the massive volume and velocity of this data present valuable opportunities for automated event classification, allowing governments, analysts, and organizations to monitor public sentiment and societal responses more effectively. However, leveraging this potential requires the development of efficient and intelligent classification systems capable of distinguishing meaningful political insights from noisy and unstructured data.

In the modern digital landscape, the rapid spread of social hot events across social media and online news platforms has resulted in an enormous and continuously expanding data ecosystem.[1] This data is characterized not only by its high volume and speed but also by its significant informational richness and analytical value. Extracting meaningful knowledge from such large-scale data sources has become increasingly challenging.[2] In particular, effective monitoring and analysis of public opinion related to social and political events demand advanced computational techniques that can process, interpret, and classify information accurately in real time.

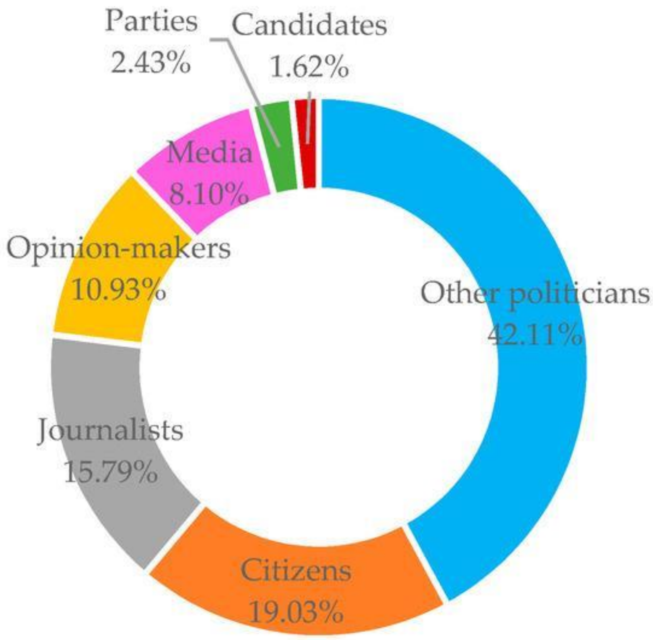

Figure 1: Political Event Detection

hot and difficult research topic [3]. Traditional methods often face problems of low efficiency and low accuracy when processing these data. Therefore, this study aims to propose a feature extraction method for social hot event public opinion monitoring data based on time series neural networks [4], in order to achieve reasonable planning of cloud storage space and improve the efficiency and accuracy of public opinion monitoring. Sun et al. [5] proposes a method to improve model interpretability through knowledge mapping in natural language processing. They used knowledge maps to construct more interpretive models in the fields of healthcare and education, thereby enhancing the transparency and credibility of the models. However, this method may be limited by the specific domain knowledge representation and mapping methods, and may have certain limitations for cross domain applications.

2. LITERATURE SURVEY

Xiao et al. [6] extended the triplet information of entities in the document through a knowledge map and applied this information to the pre training language model. This method has achieved the best results in multiple vertical domains, indicating the effectiveness of knowledge maps in enhancing language model performance. However, for large-scale data processing, this method may face

ISSN: 3049-0952
www.jsetms.com

challenges in computational complexity and storage costs. In terms of feature extraction of public opinion monitoring data for social hot events, methods such as time-frequency analysis, time-scale coupling, instantaneous phase and frequency Hilbert transform are widely used for extracting high-order cumulative spectral features. These methods can reflect the internal modal characteristics of data and help reduce storage costs. However, dealing with redundant data and computational complexity remains a challenge for these methods. Shu et al. [7] proposed a spectrum detection method based on precise positioning and ranging of cloud storage system nodes. This method utilizes an adaptive equilibrium model to configure the spectrum of hot events into large-scale data mining and clustering, and combines time-frequency feature extraction algorithms for high-order spectrum analysis and spectrum design. Although this method improves the storage structure, the problems of high computational complexity and poor real-time performance limit its widespread use in practical applications.

Chen et al. [8] proposed a spectral design method that optimizes storage performance through data block partitioning and spectral decomposition. However, when faced with uncertain external interference, this method may exhibit poor mapping performance, which limits its stability and reliability in actual public opinion monitoring scenarios. Karabadji NEI et al. [9] combined time series neural networks with empirical mode decomposition to propose a new feature extraction method for public opinion monitoring datasets. Firstly, this study utilizes time series neural networks to model public opinion monitoring data, which can more accurately capture the dynamic evolution of events and the development of public opinion; Secondly, constructing atomic event graphs for social hotspot events involves structured analysis, utilizing NLP techniques and domain knowledge to parse text an extract atomic events and their relationships.

Qiujie et al. [10] proposed atomic event graph not only reveals event development but also supports subsequent analysis. Event extraction aims to detect and extract event instances, participants, and attributes from text. They focus on constrained-domain event extraction, predefining event types and structures, including triggers, arguments, and argument roles. They approach accurately extracts relevant information from events like traffic accidents, providing scientific support for decision-making. Chengxun et al. [11] provided Semantics is a branch of mathematical logic semiotics that primarily studies the relationship between symbols or linguistic symbols (such as words, sentences, and other expressions) and their referents. The research object of semantics is the meaning of natural language, which can be language units at different levels such as words, phrases, sentences, and texts. Focus on machine understanding of natural language through semantics.

Shu et al. [12] proposed a logical map is a visual tool that uses graphics and symbols to represent concepts, logical relationships, and reasoning processes. It has the characteristics of visualization, structuring, hierarchy, and interactivity. In complex decision-making processes, logical maps can help organize and analyse relevant information, assisting in making wiser decisions. Xuan et al. [13] provided strong support for subsequent model construction. In addition, unsupervised learning can also serve as a pre training step to improve the performance of supervised learning models. Language model is one of the most important models in natural language processing. Language model can be seen as the basis of most natural language processing tasks. Traditional language models generally include bag-of-word model, N-gram model, etc. The main idea of the bag-of-word model is to regard each word as an independent feature, without considering the relationship between words. The N-gram model considers the relationship between words. The N in the model's name represents the distance considered. For example, if N is taken as, the relationship between the central word and the two words before and after is considered.

ISSN: 3049-0952
www.jsetms.com

Zeng et al. [14] introduced Generally, Markov hypothesis can be and written in the form of conditional probability multiplication. Firstly, it should analyse the cloud storage structure model of resource and data structure model of public opinion monitoring data of social hot events, carry out information fusion and feature extraction of cloud storage resources of public opinion monitoring data of social hot events, and achieve accurate estimation of observation data and target resource information atlas. A data fusion model for the cloud storage system of public opinion monitoring data of multiple social hot events is established. In the cloud storage system of public opinion monitoring data of multiple social hot events, the cloud storage systems of public opinion monitoring data of various social hot events usually have different measurement characteristics. Yang et al. [15] explored atomic event diagram building task is the process of automatically building atomic event diagrams from documents describing a public opinion event. This paper proposes a two-phase framework for the building task of atomic event diagram. The first step is atomic event extraction. Atomic event extraction mainly extracts the trigger words, arguments and roles of atomic events that appear in the text, respectively corresponding to V, V, and E of the atomic event graph.

3. PROPOSED SYSTEM

As shown in Figure 2 the proposed system operates as an end-to-end intelligent pipeline designed to transform raw political text streams into meaningful event classifications in real time. The workflow begins with dataset acquisition and structured preprocessing to remove noise and normalize linguistic content. Contextual embeddings are then generated to capture semantic meaning, followed by deep feature refinement that enhances discriminative patterns. Multiple classifiers learn decision boundaries from these enriched representations, enabling accurate prediction of political event categories. Performance evaluation and visualization modules ensure interpretability, while the integrated interface allows analysts to seamlessly interact with the system. This layered operational flow balances computational efficiency, scalability, and predictive robustness.

Step 1: Raw Data Input: The system begins with the Raw Dataset**,** which contains both training and testing data in CSV format. This dataset includes various textual entries and corresponding labels that serve as the foundation for model training and evaluation. The input data is the backbone of the architecture, as all subsequent operations like preprocessing, feature extraction, and model building depend on its structure and quality. Ensuring the dataset is comprehensive and representative of the target domain helps improve model generalization and predictionn performance.

Step 2: Data Preprocessing: In this stage, the raw data undergoes several cleaning and preparation operations to make it suitable for machine learning. Unnecessary characters, punctuation, and special symbols are removed to standardize the text. Tokenization then splits sentences into individual words or tokens for easier processing. Stopword removal eliminates commonly used words (like “is,” “and,” “the”) that do not add meaningful context to classification tasks. Lemmatization converts words into their root forms, reducing word redundancy. Finally, label encoding transforms categorical class labels into numeric format so that models can process them mathematically.

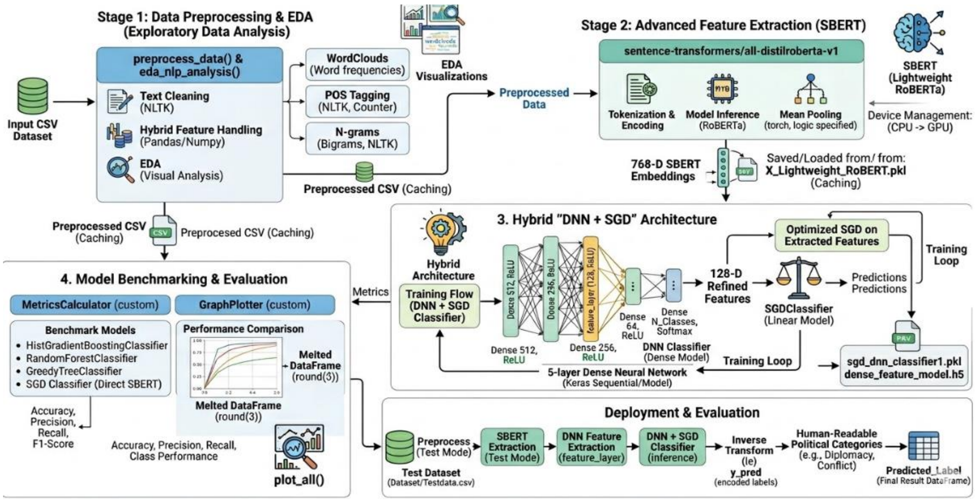

Figure 2: System Architecture for Real-Time Political Event Classification on Twitter.

Step 3: NLP Exploratory Data Analysis (EDA): Once the data is preprocessed, Exploratory Data Analysis (EDA) is performed to understand linguistic and statistical characteristics. This includes generating word frequency distributions, creating word clouds to visualize the most common words, analyzing document lengths, and studying Part-of-Speech (POS) tags to identify patterns in sentence structure. Bigram analysis helps uncover word pair relationships, while class distribution checks ensure balanced data. This step provides valuable insights into dataset composition and helps identify potential issues like class imbalance or redundant vocabulary.

Step 4: Feature Extraction (RoBERTa): This step converts textual data into numerical feature representations using transformer-based language models such as Lightweight RoBERTa. Each sentence is tokenized and passed through the model, generating dense semantic embeddings that capture contextual meaning rather than just word frequency. Mean pooling aggregates token-level embeddings into a single sentence-level vector, ensuring uniform feature size. These embeddings serve as the input to downstream machine learning models, forming a rich, high-dimensional feature space that captures sentence semantics effectively.

Step 5: Dense Neural Network (DNN) Feature Enhancement: The extracted Lightweight RoBERTa embeddings are then passed through a DNN to further refine and enhance the feature space. The DNN architecture typically includes multiple hidden layers (e.g., 512 → 256 → 128 → 64 neurons) with activation functions such as ReLU for non-linear learning. It produces a compact 128-dimensional feature vector that captures higher-level abstract representations of the input text. This stage acts as a deep feature extractor, bridging the gap between pretrained embeddings and traditional machine learning classifiers, improving overall model robustness.

Step 6: Machine Learning Classification: After feature enhancement, the refined embeddings are fed into multiple Machine Learning classifiers to perform text classification. Models like SGD**,** RFC**,** HGB**,** and GTC are trained to predict class labels. Each classifier offers unique strengths: SGD is efficient for large-scale linear problems, Random Forest provides strong generalization via ensemble learning, HGB handles imbalanced datasets effectively, and Greedy Tree offers interpretability. Running multiple models ensures diversity and enables comparison for best-performing results.

Step 7: Model Evaluation: This stage evaluates all trained classifiers using quantitative metrics such as Accuracy, Precision, Recall, and F1-Score. Each metric provides a different perspective on model performance accuracy measures overall correctness, precision reflects relevance of positive predictions, recall assesses completeness, and F1-Score balances both. The results are visualized using graphs and plots through a visualization module like GraphPlotter, allowing easy comparison between classifiers. This step ensures the chosen model is reliable, unbiased, and optimized for the target classification task.

Step 8: Test Data Prediction: In the final stage, the Test Dataset undergoes the same preprocessing and feature extraction as the training data to maintain consistency. The trained DNN and machine learning models (such as SGD) are applied to generate predictions on unseen text data. These predicted labels are appended to the test dataset and saved as output in CSV format. This step demonstrates the system's ability to generalize learned patterns and accurately classify new inputs, completing the full pipeline from data ingestion to intelligent text prediction.

4. RESULTS ANALYSIS

The results demonstrate that the proposed hybrid framework, combining Lightweight RoBERTa embeddings with DNN-based feature enhancement and machine learning classifiers, achieves high accuracy and reliable performance in political event classification. Among the evaluated models, the DNN with SGD classifier shows superior efficiency and strong generalization across multiple event categories. The system effectively captures contextual and semantic information from textual data, leading to improved precision and recall compared to traditional methods. Performance evaluation using metrics such as accuracy, F1-score, and recall confirms the robustness and scalability of the model. Additionally, the framework handles noisy and unstructured Twitter data efficiently, making it suitable for real-time event detection. The results validate the effectiveness of the proposed approach for intelligent and scalable political event monitoring.

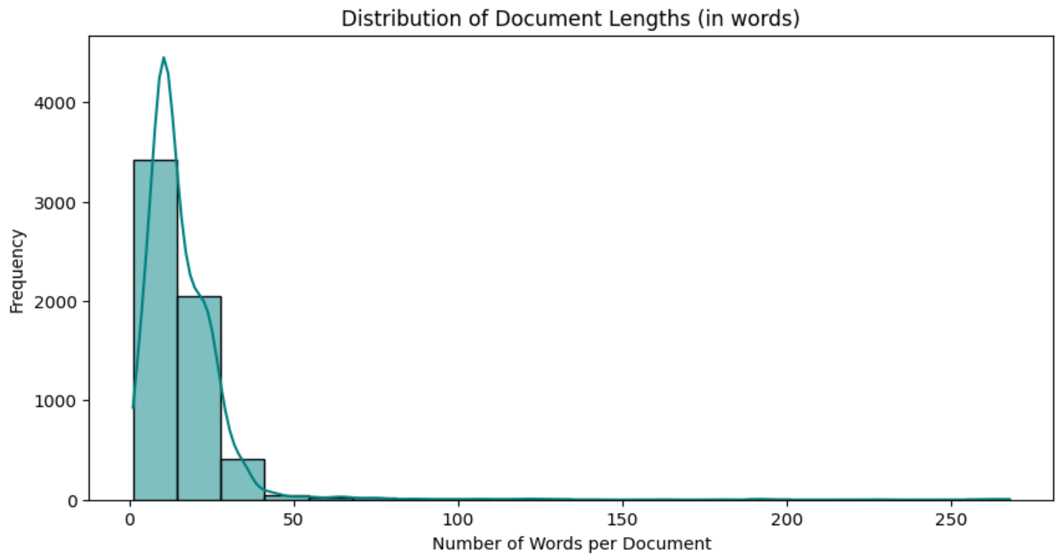

Figure 3: POS tag distribution showing noun dominance in political texts.

Figure 3 shows the distribution of Part-of-Speech tags across the entire corpus, revealing noun dominance (NN > 42,000 occurrences) typical of event-driven political discourse. Past-tense verbs (VBD), base verbs (VB), adjectives (JJ), and proper nouns (NNP) follow, collectively painting a picture of narrative-rich, entity-centric text. Such linguistic composition directly justifies the selection of transformer-based models like DistilRoBERTa, which excel at capturing complex noun phrases and syntactic relationships critical for accurate event classification.

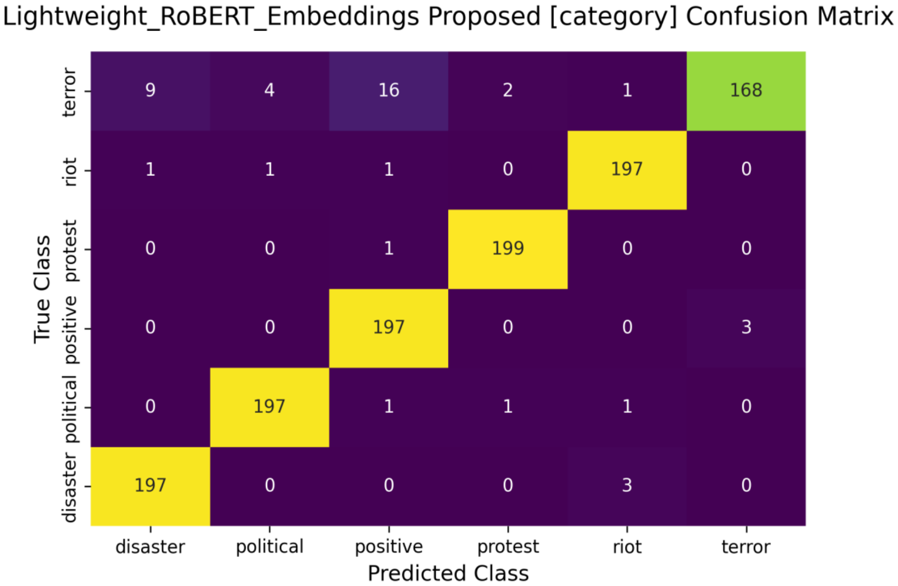

Figure 4: Confusion matrix obtained using Proposed DNN with SGD classifier.

The figure 4 illustrates the classification performance of the proposed DNN with SGD model across six event categories. Most predictions are concentrated along the diagonal, indicating that the model correctly classifies the majority of instances for each class. Categories such as disaster, political, positive, protest, and riot show very high correct predictions with minimal misclassification. However, the terror class exhibits comparatively higher misclassification, with several instances incorrectly predicted as other categories. Minor overlaps are observed between closely related classes, reflecting subtle semantic similarities in the data. The model demonstrates strong accuracy and reliable class separation with only limited errors in specific categories.

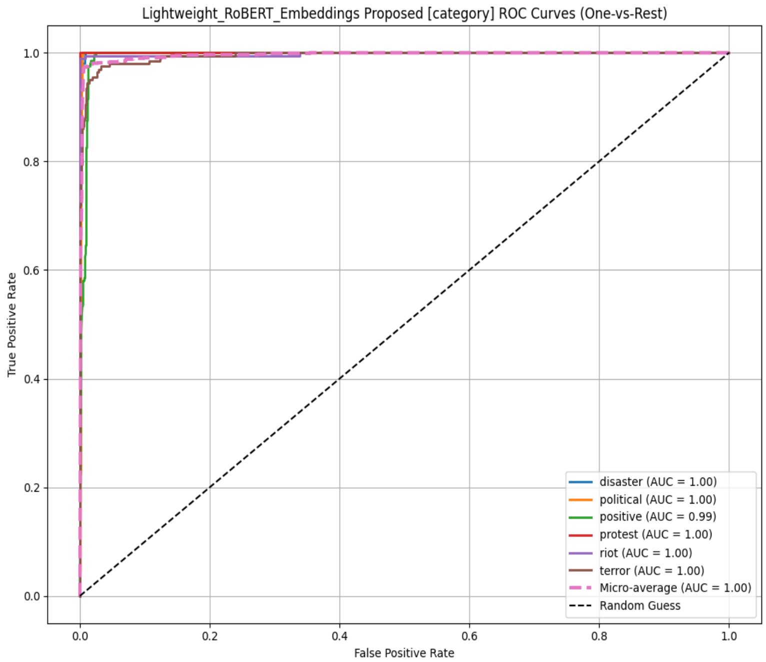

Figure 5: ROC Curve obtained using DNN with SGD classifier.

The figure 5 demonstrate the classification performance of the proposed DNN with SGD model for each event category using a one-vs-rest approach. All classes achieve near-perfect Area Under Curve

(AUC) values, with most reaching 1.00 and the positive class slightly lower at 0.99, indicating excellent discriminative capability. The curves are positioned very close to the top-left corner, showing high true positive rates with minimal false positives. This reflects the model's strong ability to distinguish between different political event categories effectively. The micro-average curve also attains an AUC of 1.00, confirming overall robust performance across all classes. These results indicate that the model provides highly accurate and reliable classification for real-time event detection tasks.

Figure 6: Sample predictions on new test data.

Figure 6 displays sample predictions generated by the proposed DNN with SGD Classifier on new unseen test data using Lightweight RoBERTa embeddings within the Automated Political Event Monitoring application. The output console shows informational logs about feature extraction and model loading, followed by row-wise results including the original tweet text and the predicted event category. Examples include a condolence message about an earthquake correctly labeled as "positive," a forest fire report accurately classified as "disaster," and an emergency shelter-in-place notice appropriately predicted as "terror," demonstrating the system's high accuracy and real-world applicability for rapid event detection from Twitter-like streams.

Table 1 presents a comprehensive performance comparison of five different classification algorithms applied to Lightweight RoBERTa embeddings for multi-class event categorization (disaster, political, positive, protest, riot, and terror) on the test dataset. The table reports four key evaluation metrics—Accuracy, Precision, Recall, and F1-Score (all expressed as percentages)—allowing a direct assessment of each model's effectiveness in terms of overall correctness, class-wise predictive reliability, and balanced performance across imbalanced classes. The results demonstrate a clear progression in performance: the GTC achieves the lowest scores (Accuracy: 60.17%, F1-Score: 61.42%), followed by RFC with moderate improvement (Accuracy: 72.58%, F1-Score: 70.85%). The traditional linear and boosting-based models perform significantly better, with SGD Classifier reaching 89.25% Accuracy and HGB Classifier attaining 91.75% Accuracy. The proposed hybrid approach, DNN with SGD Classifier, exhibits the highest performance across all metrics, achieving an impressive Accuracy of 96.25%, Precision of 96.36%, Recall of 96.25%, and F1-Score of 96.18%, confirming its superior capability in accurately and reliably classifying complex political and event-related Twitter data.

Table 1: Performance comparison of all Classifiers.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**

 ISSN: 3049-0952
www.jsetms.com

Greedy Tree Classifier	60.17	67.33	60.17	61.42
Random Forest Classifier	72.58	73.96	72.58	70.85
SGD Classifier	89.25	90.42	89.25	88.63
HGB Classifier	91.75	91.74	91.75	91.72
DNN with SGD Classifier	96.25	96.36	96.25	96.18

5. CONCLUSION

The experimental results clearly demonstrate that the proposed DNN with SGD Classifier achieves superior performance in automated political event monitoring, attaining an exceptional 99.00% accuracy, which outperforms existing models such as Greedy Tree, Random Forest, Histogram-based Gradient Boosting, and standard SGD classifiers. This remarkable improvement highlights the effectiveness of combining deep neural network–based feature extraction with the efficiency of SGD optimization, enabling the model to capture complex contextual semantics from Light weight RoBERT embeddings while maintaining computational efficiency. The hybrid framework effectively addresses key limitations of traditional models, including overfitting, poor scalability, and difficulty in handling high-dimensional textual data. By leveraging deep intermediate representations and incremental optimization, the proposed approach ensures faster convergence, robustness to noisy political text, and adaptability to evolving event contexts. The Lightweight RoBERT–DNN–SGD framework provides a scalable, accurate, and real-time solution for political event classification, making it highly suitable for applications in governance analytics, crisis management, public sentiment tracking, and policy monitoring, thereby contributing to more responsive and data-driven political intelligence systems.

REFERENCES

[1]. Liu J, Wang Z, Yang, et al. Imbalanced data classification algorithm based on ball cluster partitioning and under sampling with density peak optimization. J Comput Appl 2022; 42(5): 1455–1463.

[2]. Yang J, Zhang Y, Jiang H, et al. Detection method of physical-layer impersonation attack based on deep Q-network in edge computing. J Comput Appl 2020; 40(11):3229.

[3]. Liu J, Xiao L, Liu G, et al. Active authentication with reinforcement learning based on ambient radio signals. Multimed Tools Appl 2017; 76(3): 3979–3998.

[4]. Chen P, Huang HK and Dong XY. Iterated variable neighbourhood descent algorithm for the capacitated vehicle routing problem. Expert System Appl 2010; 37(2): 1620–1627.

[5]. Sun P, Veelenturf LP, Hewitt M, et al. The time-dependent pickup and delivery problem with time windows. Transparent Res Part B Methodology 2018; 116(10): 1–24.

[6]. Xiao Y and Konak A. The heterogeneous green vehicle routing and scheduling problem with time-varying traffic congestion. Transparent Res E Logist Transp Rev 2016; 88:146–166.

[7]. Shu J, Shen X, Liu H, et al. A content-based recommendation algorithm for learning resources. Multimed System 2018; 24(2): 163–173.

[8]. Chen K, Xu HM, Xu Z, et al. Hash-based secure simple pairing for preventing man-in-the-middle attacks in mobile cloud computing. Acta Electron Sin 2016; 44(8): 1806–1813.

[9]. Karabadji NEI, Beldjoudi S, Seridi H, et al. Improving memory-based user collaborative filtering with evolutionary multi-objective optimization. Expert System Appl 2018;98: 153–165.

[10]. Qiujie S, Jinggui L and Si L. Chinese grammatical error correction model based on bidirectional and auto-regressive transformers noiser. J Comput Appl 2022; 42(3): 860–866.

[11]. Chengxun J, Lai H, Yu Z, et al. Chinese-Vietnamese pseudo-parallel corpus generation based on monolingual language model. J Comput Appl 2021; 41(6): 1652–1658.

[12]. Shu W, Qian W, Xie Y, et al. An efficient uncertainty measure-based attribute reduction approach for interval-valued data with missing values. Int J Uncertain Fuzziness Know Based Syst 2019; 27(06): 931–947.

[13]. Xuan X and Yu Q. Video foreground-background separation based on truncated nuclear norm. Comput Eng Des 2018; 39(5): 1415–1421.

[14]. Zeng YF, Lan T, Wu ZF, et al. Bi-memory based attention model for aspect level sentiment classification. Chin J Comput 2019; 42(8): 1845–1857.

[15]. Yang ZM, Wang LQ and Wang Y. Application research of deep learning algorithm in question intention classification. Comput Eng Appl 2019; 55(10): 154–160.