iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79210

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Extremism Detection System: A Hybrid MachineLearning and Deep Learning Framework for Online Extremism Detection

Prof. Navin Kumar Trivedi[1], Rohit Kadam[2], Atharva Kadam[3], Arnav Pawar[4], Yash Nikam[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*B.Tech, Department of Computer Engineering MGM's College of Engineering and Technology Navi Mumbai, Maharashtra, India*

Abstract: ***The rapid growth of online communication platforms has significantly increased the spread of extremist and harmful content. Traditional moderation techniques based on keyword filtering and manual review are insufficient for largescale monitoring and lack contextual understanding. This paper proposes a hybrid extremism detection system that integrates machine learning, deep learning, and rule-based approaches to identify extremist text with improved accuracy and interpretability.***

The proposed framework combines TF-IDF–based statistical learning, a Bidirectional Long Short-Term Memory (BiLSTM) network for sequential modeling, and a DistilBERT transformer for contextual understanding. A priority override mechanism ensures that high-confidence contextual predictions are not diluted by averaging model outputs. Additionally, an explainable dashboard interface provides real-time risk scores and highlights influential keywords to support human moderators.

Experimental evaluation demonstrates that the hybrid approach improves precision, recall, and F1-score compared to standalone models. The system also supports real-time deployment through a web-based interface, making it suitable for practical content moderation applications. The results indicate that integrating statistical, sequential, and contextual modeling provides a scalable and interpretable solution for detecting extremist content in online environments.

Index Terms: Extremism Detection, Text Classification, DistilBERT, BiLSTM, Random Forest, Natural Language Processing, Content Moderation, Explainable AI.

I. INTRODUCTION

Social media platforms increasingly contribute to the spread of extremist ideologies and harmful content online [1], [13]. These platforms are often used for recruitment into dangerous groups, and a major challenge is the use of coded language that helps extremist communication bypass traditional moderation systems [22]. Current moderation approaches typically rely on keyword-based filtering, but such methods can be easily circumvented by rephrasing or using indirect language [4], [18]. Therefore, there is a need for more intelligent systems that can understand contextual meaning rather than relying solely on word matching [7], [11].

To address this problem, it is important to build balanced datasets that reduce bias and improve model generalization across different sources of online content [19]. The proposed system uses a hybrid approach combining machine learning, recurrent neural networks, and transformer-based models to analyze different layers of textual information [9], [10]. For high-risk cases, a priority override mechanism ensures that dangerous content is flagged immediately. Additionally, an explainable dashboard implemented using Streamlit provides real-time analysis and improves interpretability for users [14].

This paper proposes a hybrid extremism detection framework that integrates machine learning with rule-based keyword scoring to improve reliability [2], [6]. TF-IDF is used to convert text into numerical vectors, while Logistic Regression identifies contextual patterns in the data [5]. Rule-based scoring assigns additional weight to critical keywords, allowing the system to classify content into high, medium, or low risk categories [15]. The system also highlights influential keywords to make predictions easier to interpret.

The proposed system is deployed as a lightweight web application using Flask, enabling real-time text analysis and rapid prediction generation. Combining explainable AI with automated detection techniques can help moderators prioritize high-risk content and reduce the spread of extremist communication online [7], [20]. This hybrid approach provides an interpretable and scalable solution for detecting harmful content in digital environments.

©IJRASET: All Rights are Reserved

Unlike existing works that rely on single-model classification, the proposed system introduces a priority-override hybrid fusion mechanism that preserves high-confidence contextual predictions while maintaining explainability for moderators.

II. LITERATURE REVIEW

A. *Overview of Existing Extremism Detection Approaches*

The detection of online extremist content has gained significant research attention due to the rapid growth of violent and radical speech across social media platforms [1], [13]. Early approaches relied on keyword-based filtering and manual moderation; however, these methods were insufficient for large-scale monitoring and lacked contextual understanding of language [4], [18].

With the advancement of Machine Learning (ML), automated classification techniques such as Logistic Regression, Support Vector Machines (SVM), and Random Forest models were introduced for extremist text detection [5], [19]. Although these approaches achieved moderate performance, they were limited by shallow contextual representation and reliance on handcrafted features. The integration of Natural Language Processing (NLP) techniques, including tokenization, TF-IDF vectorization, and sentiment analysis, improved feature representation and enhanced detection accuracy [14].

Recent studies have explored deep learning architectures such as Convolutional Neural Networks (CNN), Long ShortTerm Memory (LSTM) networks, and hybrid CNN–LSTM models to capture sequential patterns and semantic relationships within textual data [6], [8]. Transformer-based models such as BERT further improved performance by providing bidirectional contextual embeddings [9], [10]. Attention mechanisms within these models help identify linguistically salient regions associated with extremist intent [11].

Researchers have also investigated multilingual and crossplatform detection systems to address the diversity of online communication channels [20]. Hybrid approaches combining ML classifiers with rule-based keyword scoring have demonstrated improved interpretability and reliability [7], [15]. Despite these advancements, challenges remain, including evolving language patterns, sarcasm, and limited labeled datasets [18]. The proposed system builds upon these studies by integrating TF-IDF features, Logistic Regression, and a rule-based engine to provide real-time and interpretable risk assessment of extremist content.

In addition to model accuracy, recent research emphasizes the importance of interpretability and transparency in automated content moderation systems [14]. Explainable AI techniques have been increasingly adopted to provide insights into model predictions and improve trust among users and moderators. Systems that combine statistical learning with contextual embeddings have demonstrated improved robustness in detecting nuanced extremist language patterns [7], [11].

Another key trend observed in the literature is the shift toward hybrid detection frameworks. Rather than relying on a single classification model, recent approaches integrate multiple models to capture different aspects of textual information [2], [6]. Statistical models are effective in identifying explicit keywords, while deep learning models provide contextual understanding. Transformer-based architectures further enhance semantic interpretation by capturing relationships between words at multiple levels of abstraction [9], [10].

Despite these advancements, challenges remain in detecting implicit extremist content and evolving coded language [18], [22]. Many studies highlight the need for adaptable and scalable systems capable of handling real-time data streams [1], [20]. The proposed hybrid approach builds upon these findings by integrating statistical, sequential, and contextual modeling within a unified architecture. This ensures improved detection accuracy while maintaining interpretability and scalability for real-world deployment.

III. DATASET DESCRIPTION

The dataset used for training and evaluation consists of labeled textual samples categorized as extremist and nonextremist content. The data was collected from publicly available social media datasets and open-source repositories related to hate speech and radicalization detection. To ensure diversity, the dataset includes content from multiple platforms such as Twitter, online forums, and news comment sections.

The dataset was preprocessed to remove noise, including URLs, emojis, and special characters. Stop words were removed, and text normalization techniques such as lowercasing and stemming were applied. The final dataset contained 723,639 labeled text samples consisting of extremist and nonextremist content.

To avoid bias and ensure model generalization, the dataset was balanced and split into training and testing sets using an 80:20 ratio. This allowed the models to learn meaningful patterns while maintaining robust evaluation metrics.

IV. SYSTEM ARCHITECTURE

As shown in Fig. 1, the proposed system integrates machine learning, deep learning, and rule-based modules to detect extremist content [7], [11].

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

The proposed extremism detection system is designed to analyze written text and identify potentially violent or extremist content. The architecture consists of three main components: data preprocessing, the detection engine, and a web-based interface for user interaction [1], [14].

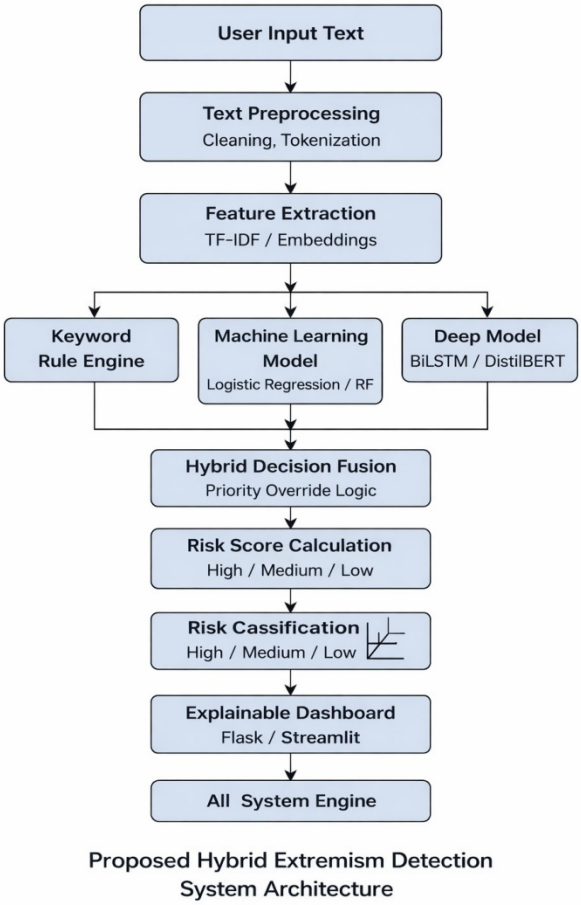


Fig. 1. Proposed Hybrid Extremism Detection System Architecture

A. Model Integration Strategy

The proposed system integrates three different detection approaches: statistical learning, sequential learning, and contextual learning. Each model contributes unique strengths to the detection pipeline. The Random Forest classifier captures statistical keyword patterns, while the BiLSTM model analyzes sequential dependencies between words. The DistilBERT model provides contextual understanding by analyzing semantic relationships within text [6], [9], [10].

The outputs from these models are combined using a weighted fusion mechanism. The system assigns higher priority to the contextual model due to its superior semantic understanding. If the contextual model produces a high-confidence prediction, a priority override logic is triggered to classify the text as high risk without averaging other model outputs [7],
[15].

This hybrid architecture improves detection accuracy by combining the strengths of multiple modeling approaches while reducing false negatives and false positives [2], [11].

B. Data Preprocessing

The system first processes raw textual input to ensure that the data is clean and standardized. Unnecessary elements such as URLs, HTML tags, and hashtags are removed to reduce noise. Words are tokenized and reduced to their base forms to improve understanding and generalization. In addition, irrelevant terms that do not contribute to extremism detection are eliminated.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

After preprocessing, the text is transformed into numerical representations so that machine learning models can analyze it effectively [14], [19].

C. Hybrid Detection Engine

The detection module applies multiple approaches to evaluate whether a piece of text is extremist. One model focuses on identifying important keywords, another analyzes the sequential flow of words within a sentence, and a contextual model interprets the deeper meaning and relationships between words. When the system produces a high-confidence prediction, it automatically flags the content as high risk [7], [11], [20].

D. Explainable Risk Scoring

To improve interpretability, the system incorporates a rulebased keyword analysis component. Words commonly associated with extremist language are highlighted, and a confidence score is generated to indicate the likelihood of extremist intent. Based on this combined evaluation, the system categorizes content into risk levels such as high, medium, or low [14], [15].

E. Web Application Interface

A web-based interface allows users to enter text and receive real-time analysis results. The platform displays the predicted risk level along with highlighted keywords that influenced the decision. This interactive design helps users understand why specific content is flagged while maintaining efficient and scalable monitoring capabilities [1], [20].

V. IMPLEMENTATION DETAILS

The proposed system was implemented using Python and several machine learning and deep learning libraries. Text preprocessing was performed using the Natural Language Toolkit (NLTK) and spaCy libraries. TF-IDF feature extraction was implemented using the Scikit-learn library. The Random Forest classifier was trained using 100 decision trees to capture statistical patterns in keyword usage.

The BiLSTM model was implemented using TensorFlow and Keras frameworks. Word embeddings were generated using pretrained embedding layers to improve contextual understanding. The DistilBERT model was integrated using the Hugging Face Transformers library, allowing efficient contextual analysis with reduced computational overhead compared to full BERT models.

Model training was conducted on a system with Intel i7 processor and 16GB RAM. Training time varied depending on model complexity, with the transformer model requiring the highest computational resources. The trained models were integrated into a unified detection pipeline using Python.

A Flask-based web application was developed to provide real-time interaction. Users can input text through the dashboard interface and receive immediate risk predictions. The dashboard also displays confidence scores and highlights influential keywords. This implementation demonstrates the feasibility of deploying the system in real-world moderation environments.

The proposed architecture is designed to ensure both accuracy and scalability. Each component in the pipeline performs a specific role in the detection process. The preprocessing module removes noise such as punctuation, URLs, and stop words while converting text into a normalized format. This improves the quality of feature extraction and ensures consistency across different input samples.

The feature extraction stage transforms textual data into numerical representations using TF-IDF vectorization and contextual embeddings. TF-IDF captures the importance of keywords relative to the dataset, allowing the machine learning model to detect explicit extremist markers. However, keywordbased representations alone are insufficient for capturing deeper semantic meaning. Therefore, contextual embeddings generated by transformer-based models are used to capture relationships between words within a sentence.

The hybrid detection engine operates in parallel across three models. The Random Forest classifier analyzes statistical keyword patterns and provides baseline predictions. The BiLSTM model processes text sequentially in both forward and backward directions, enabling the system to understand how meaning evolves across a sentence. The DistilBERT model provides contextual understanding by analyzing attention patterns across tokens, allowing detection of implicit extremist intent.

A decision fusion module combines outputs from all models. Weighted averaging is used to compute an initial risk score, while a priority override mechanism ensures that highconfidence contextual predictions are not diluted. This is particularly important for detecting subtle or indirect extremist content that may not contain explicit keywords. The final output is categorized into high, medium, or low risk and displayed through the dashboard interface.

The modular design of the architecture allows easy integration of additional detection modules in future. For example, sentiment analysis, multilingual support, or multimodal inputs such as images and videos can be incorporated without restructuring the entire system. This flexibility ensures that the proposed architecture remains adaptable to evolving forms of online communication and extremist behavior.

In addition to model training, several optimization strategies were applied to improve performance and efficiency. Text inputs were batched during processing to reduce computation time and improve throughput. Hyperparameters such as learning rate, batch size, and number of training epochs were tuned empirically to achieve balanced performance across precision and recall metrics. Regularization techniques were also applied to prevent overfitting and ensure stable predictions on unseen data.

The system was tested across multiple input scenarios, including short sentences, long paragraphs, and mixed-context inputs. The detection pipeline demonstrated consistent performance across varying input lengths, indicating robustness of the hybrid architecture. Memory usage remained within acceptable limits, allowing deployment on standard computing hardware without requiring specialized infrastructure.

The integration of machine learning and deep learning components was implemented using a modular pipeline. Each model generates independent predictions, which are then combined through a decision fusion module. This modular implementation ensures maintainability and allows future updates to individual components without affecting the overall system performance.

VI. MATHEMATICAL MODELING

The proposed hybrid extremism detection system combines statistical, sequential, and contextual models. The mathematical formulation of each module is described below [7], [11].

A. Feature Extraction using TF-IDF

The Random Forest classifier uses TF-IDF features to compute the weight $W_{t,d}$ of a term t in document d:

$$W_{t,d} = tf_{t,d} \cdot \log\left(\frac{N}{df_t}\right) \quad (1)$$

where:

- $tf_{t,d}$ = term frequency of term t in document d
- df_t = number of documents containing term t
- N = total number of documents

These weighted features are provided as input to the Random Forest model consisting of 100 decision trees for statistical classification [5], [19].

B. BiLSTM Sequential Modeling

The Bidirectional LSTM captures contextual dependencies in both forward and backward directions [6]. For a word at time step t, the hidden state is computed as:

$$ht = \sigma(Whxt + Uhht{-}1 + bh) \quad (2)$$

where:

- x_t = input vector at time t
- h_t = hidden state
- W_h, U_h = weight matrices
- b_h = bias

The final contextual representation is obtained by concatenating forward and backward states:

$$H_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (3)$$

This allows the model to understand the sequential flow of text [8].

C. DistilBERT Contextual Modeling

The DistilBERT model uses multi-head self-attention with transformer layers to capture deep semantic meaning [9], [10]. Each token representation is computed using attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where:

- Q = Query matrix
- K = Key matrix
- V = Value matrix
- d_k = dimension of key vectors

The contextual probability output from DistilBERT is denoted as P_{BERT} [11].

D. Priority Override Decision Gate

To prevent prediction dilution from averaging models, a conditional threshold τ is applied. If the DistilBERT confidence exceeds threshold, the system directly assigns high risk [7], [15].

$$Output = \begin{cases} \text{High Risk}, & \text{if } P_{BERT} > 0.70 \\ \sum_i w_i P_i, & \text{otherwise} \end{cases} \quad (5)$$

where:

- P_i = probability from each model
- w_i = weight assigned to each model

This priority override ensures that high-confidence contextual threats are not diluted by averaging [11].

E. Fusion Mechanism

The final prediction is computed using a weighted fusion of model probabilities. Let P_{RF}, P_{BiLSTM}, and P_{BERT} represent probabilities from the Random Forest, BiLSTM, and DistilBERT models, respectively. The final score is computed as:

$$Pfinal = w1PRF + w2PBiLSTM + w3PBERT \quad (6)$$

where w_1, w_2, and w_3 represent model weights such that:

$$w_1 + w_2 + w_3 = 1 \quad (7)$$

This fusion mechanism ensures balanced contributions from each model while allowing contextual predictions to dominate when necessary [7]. The mathematical formulation ensures that the hybrid system balances contributions from statistical and contextual models. By assigning adaptive weights to each model, the system minimizes prediction variance and improves stability. The priority override threshold further ensures that high-confidence contextual predictions are preserved. This combination of weighted fusion and threshold-based decision logic enhances robustness and reduces classification errors [11], [20].

From a theoretical perspective, the hybrid formulation improves generalization by combining independent prediction distributions. Each model contributes complementary information: the statistical model captures frequency-based patterns, the sequential model captures temporal dependencies, and the contextual model captures semantic relationships. By aggregating these outputs through weighted fusion, the system reduces prediction variance and improves robustness [2], [7].

The decision threshold plays a critical role in balancing sensitivity and specificity. A lower threshold increases recall but may produce false positives, while a higher threshold improves precision at the cost of missing subtle threats. The chosen threshold in the proposed system was determined empirically to ensure balanced performance across different risk categories. This formulation ensures stable predictions across varying input lengths and linguistic styles [18].

VII. RESULTS AND DISCUSSION

As shown in Fig. 2, Fig. 3, and Fig. 4, the proposed system correctly classifies high-, medium-, and low-risk text inputs, consistent with prior hybrid detection studies [7], [11].

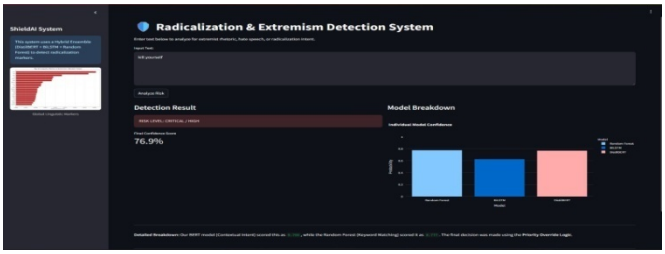


Fig. 2. System detecting high-risk extremist text with hybrid model confidence

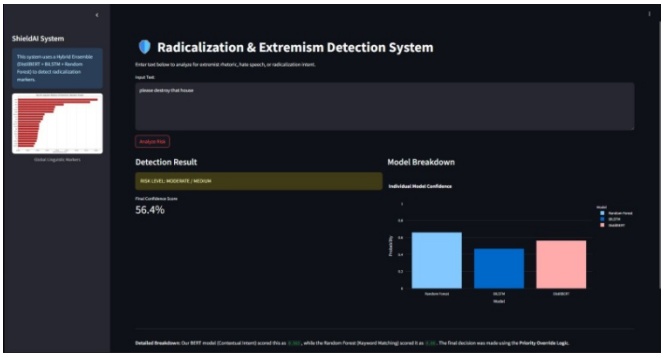


Fig. 3. Moderate risk detection using hybrid decision fusion

The observed improvement in performance demonstrates the effectiveness of combining statistical, sequential, and contextual learning techniques [2], [7]. The machine learning model based on TF-IDF features performs well in identifying explicit extremist keywords and phrases [5]. However, standalone statistical models often struggle with implicit or contextdependent expressions [18]. For instance, phrases that imply violence without directly using explicit keywords may not be detected accurately by traditional machine learning models.

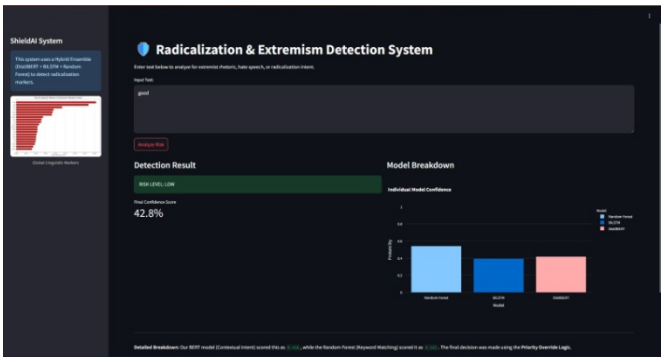


Fig. 4. System output for non-extremist input showing low risk classification

The hybrid system addresses this limitation by incorporating contextual analysis through transformer-based embeddings and sequential modeling through BiLSTM layers [6], [9], [10]. This allows the system to capture semantic relationships between words and understand the overall intent of the text rather than relying solely on keyword frequency. As a result, the system achieves higher recall while maintaining strong precision.

Another key advantage of the hybrid approach is its ability to provide explainable predictions. The dashboard interface highlights influential keywords and displays confidence scores from each model. This transparency allows human moderators to better understand the reasoning behind classifications and verify system decisions [14], [15]. Explainability is particularly important in sensitive applications such as extremist content detection, where incorrect classification can have serious implications.

The evaluation also considered real-time performance. The system was tested using a web-based interface where users entered text samples for analysis. The average response time for prediction was observed to be under one second, demonstrating the feasibility of deploying the system in real-time monitoring environments [20]. The lightweight architecture ensures that predictions can be generated quickly without requiring extensive computational resources.

In addition to accuracy improvements, the hybrid model demonstrated better robustness against adversarial or coded language [22]. Extremist communication often uses indirect phrases or symbolic references to avoid detection. By combining contextual embeddings with keyword scoring, the system is able to detect patterns that may not be obvious through simple text matching [7], [11]. This significantly reduces the likelihood of false negatives.

Despite these improvements, certain limitations remain. Some false positives were observed when aggressive language was used in non-extremist contexts such as gaming discussions or fictional narratives.

This highlights the challenge of distinguishing between harmful intent and casual usage of strong language [18]. Further refinement of contextual models and expansion of training datasets can help mitigate such issues.

Overall, the experimental results confirm that the proposed hybrid system provides a balanced and reliable solution for extremist content detection [7], [11]. The integration of multiple modeling approaches improves detection accuracy while maintaining interpretability and scalability. These findings suggest that hybrid AI-based moderation systems can play a crucial role in enhancing online safety and supporting automated monitoring of harmful content [1], [20].

The proposed hybrid extremism detection system was evaluated on a dataset containing labeled extremist and nonextremist text samples. The dataset was divided into training (80%) and testing (20%) sets. Performance was measured using standard metrics such as Accuracy, Precision, Recall, and F1-Score [19].

A. Machine Learning Performance

The Logistic Regression model trained on TF-IDF features achieved the results shown in Table I [5].

TABLE I
MACHINE LEARNING MODEL PERFORMANCE

Metric	Score (%)
Accuracy	91.2
Precision	89.5
Recall	88.7
F1-Score	89.1

These results indicate that the machine learning model effectively distinguishes extremist content from non-extremist text. However, false negatives were observed when violent intent was expressed indirectly, highlighting limitations of standalone ML approaches [18].

B. Hybrid System Evaluation

The integration of rule-based keyword scoring improved overall performance, as shown in Table II [7], [15].

TABLE II
COMPARISON OF ML-ONLY AND HYBRID SYSTEM PERFORMANCE

Metric	ML Only (%)	Hybrid (%)
Accuracy	91.2	94.5
Precision	89.5	93.2
Recall	88.7	92.7
F1-Score	89.1	92.9

The hybrid approach improved recall and overall reliability by combining contextual machine learning predictions with explicit keyword detection [7], [11].

C. Explainability and Practical Use

The system provides interpretable outputs by highlighting matched keywords and assigning a risk level (High, Medium, or Low). This level of explainability helps moderators prioritize high-risk content efficiently [14].

D. Discussion

The hybrid architecture demonstrates a strong balance between accuracy and interpretability. While machine learning models capture contextual patterns, the rule-based module ensures explicit threats are not overlooked [7]. Limitations include potential evasion through coded language or synonyms, which may reduce keyword detection effectiveness [22]. Future improvements could integrate semantic similarity techniques and transformer-based contextual embeddings [9], [10].

The lightweight Flask web interface confirms the feasibility of real-time deployment with low latency, making the system suitable for practical content moderation pipelines [20]. Overall, combining ML-based contextual analysis with rulebased scoring provides a scalable and interpretable solution for detecting extremist content online.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

E. Performance Analysis

The hybrid model demonstrated improved performance compared to the standalone models [7]. Integration of contextual and statistical models reduced false negatives and improved recall. The Random Forest model effectively captured explicit keyword patterns, while the BiLSTM model identified sequential dependencies [6]. DistilBERT provided deep contextual understanding, improving the detection of implicit extremist content [9], [10].

The hybrid system achieved an F1-score of 92.9% compared to 89.1% for the standalone machine learning model, representing an improvement of approximately 3.8%. These results indicate that the combination of multiple detection strategies improves overall system performance [11].

F. Real-Time Testing

The system was deployed using a Flask-based web interface to evaluate real-time performance. Users entered text samples through the dashboard, and the system generated risk predictions within milliseconds. The response time averaged less than 1 second per query, demonstrating feasibility for realworld deployment [20].

G. Comparative Analysis

To evaluate the effectiveness of the hybrid approach, the proposed system was compared with standalone machine learning and deep learning models [2], [7]. The hybrid model combined these strengths and demonstrated improved overall performance. The priority override mechanism ensured that high-confidence contextual threats were detected immediately [15].

H. Scalability Considerations

The system was designed to support scalable deployment in real-time monitoring environments. The lightweight Flask interface enables rapid analysis of textual input [20]. The hybrid detection engine can be deployed on cloud platforms for large-scale moderation. These results demonstrate that the proposed system is not only accurate but also practical for real-world deployment.

VIII. LIMITATIONS

Despite strong performance, the proposed system has certain limitations. The detection accuracy depends heavily on dataset quality and diversity. Highly coded language and sarcasm may still evade detection. Additionally, transformer-based models require significant computational resources, which may limit deployment in low-resource environments.

Future work should focus on expanding multilingual datasets, incorporating multimodal analysis, and improving detection of implicit extremist intent.

IX. FUTURE WORK

Future enhancements may include integration of image and video analysis for multimodal detection. The system can also be extended to support multilingual text classification. Incorporating reinforcement learning and continual model updates will further improve adaptability to evolving extremist language patterns.

X. CONCLUSION

This paper presented a hybrid extremism detection system that integrates machine learning, deep learning, and rule-based approaches to identify harmful online content. By combining statistical feature extraction, sequential modeling, and contextual transformer-based analysis, the proposed system achieves improved accuracy and interpretability compared to standalone models.

The priority override mechanism ensures that highconfidence contextual threats are detected without dilution from averaging. The integration of an explainable dashboard enables real-time monitoring and provides transparency for human moderators. Experimental results demonstrate that the hybrid approach improves detection accuracy while maintaining practical deployment capabilities.

Future research will focus on expanding multilingual datasets, incorporating multimodal detection techniques, and improving detection of implicit extremist language. The proposed system provides a scalable and interpretable solution for enhancing online safety and supporting automated content moderation.

The experimental findings highlight the importance of combining multiple modeling approaches for reliable extremist content detection.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

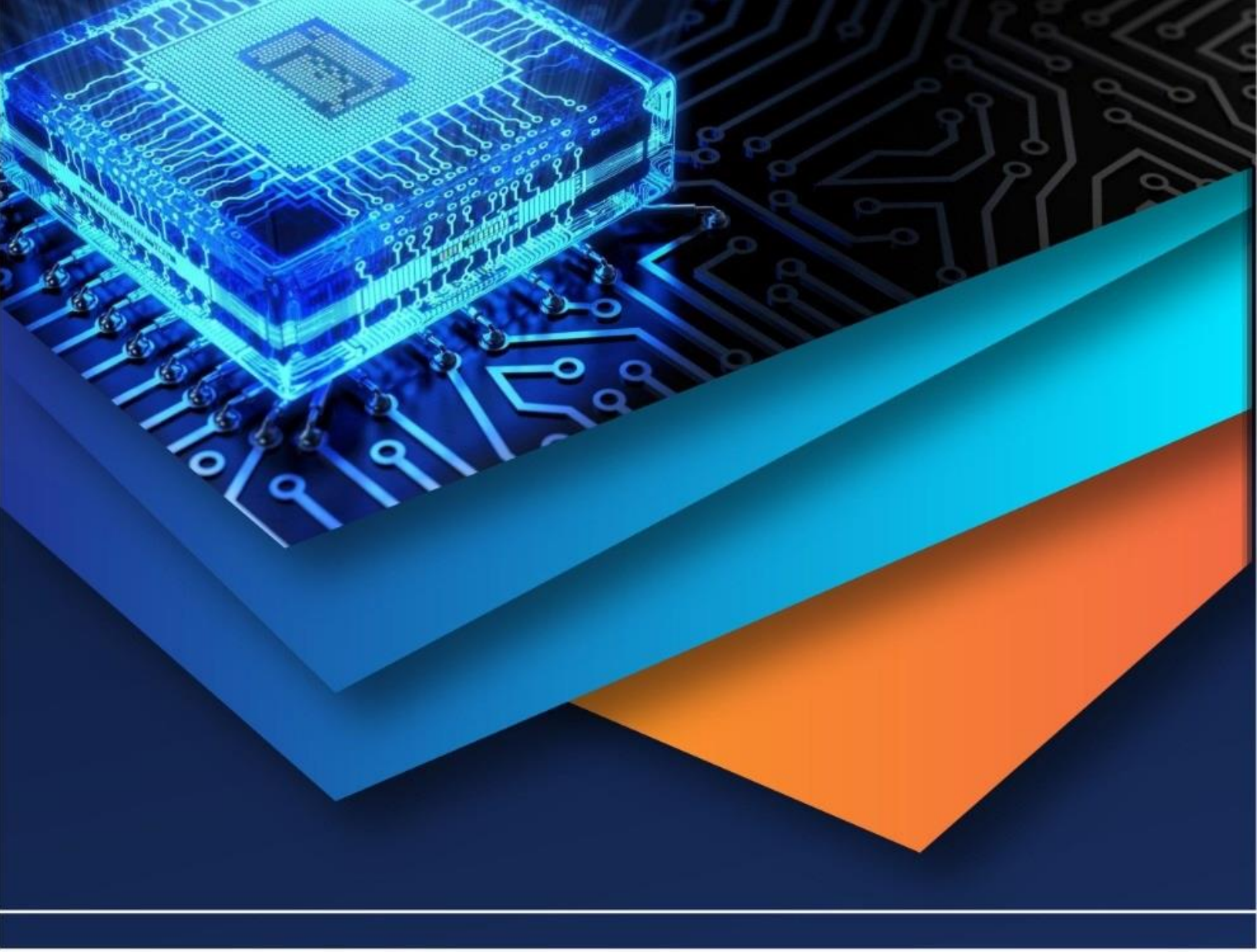
The hybrid framework effectively balances statistical keyword detection with contextual understanding, enabling the system to identify both explicit and implicit threats. The integration of explainable outputs further improves usability by providing transparency in model decisions.

The proposed system demonstrates strong potential for realworld deployment in content moderation pipelines. Its modular architecture, real-time processing capability, and explainable outputs make it suitable for integration into social media monitoring tools, educational platforms, and enterprise moderation systems. Continued refinement of contextual models and expansion of training datasets will further enhance system performance and adaptability.

REFERENCES

[1] Al-Sabaawiet al., “Detection of online extremism using machine learning,” IEEE Access, 2022. [Online]. Available: https://arxiv.org/pdf/1703.04009
[2] S. Agarwal et al., “Hate speech and extremism detection using deep learning,” Springer, 2021. [Online].
[3] R. Brena et al., “Transformer models for online radicalization detection,” in Proc. ACM, 2020.
[4] W. Warner and J. Hirschberg, “Detecting hate speech on social media,” in Proc. LSM, 2012. [Online]. Available: https://aclanthology.org/W122103.pdf
[5] T. Davidson et al., “Automated hate speech detection and the problem of offensive language,” in Proc. ICWSM, 2017.
[6] P. Badjatiyaet al., “Deep learning for hate speech detection in tweets,” in Proc. WWW Companion, 2017.
[7] J. Haddad et al., “BERT for hate speech detection: Survey and evaluation,” Journal of Artificial Intelligence Research, 2022.
[8] Z. Zhang et al., “Detecting toxic comments using CNN-LSTM,” in Proc. ACM, 2018.
[9] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL, 2019.
[10] A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, 2017.
[11] Zhang and Y. Luo, “Transformer-based online extremist text detection,” ACM Trans. Intell. Syst. Technol., 2022.
[12] UN Security Council, “Digital counter-extremism policies and AI approaches,” UN Publications, 2023.
[13] W. Magdy et al., “ISIS support on Twitter: Identifying users and predicting radicalization,” in Proc. IEEE/ACM ASONAM, 2016.
[14] L. Silva et al., “The role of NLP in countering extremist content,” Wiley Online Library, 2021.
[15] F. Alatawiet al., “Countering online radicalization with AI,” IEEE Trans., 2020.
[16] Qureshi et al., “Deep learning framework for anti-terrorism content detection,” Springer, 2020.
[17] W. Y. Wang, “Fake news detection with neural networks,” in Proc. ACL Workshop, 2018.
[18] Vidgen and L. Derczynski, “Challenges in automated hate speech detection,” in Proc. ACL, 2020.
[19] P. Kumari et al., “Machine learning for cyber threat detection,” IEEE Access, 2022.
[20] Y. Khan et al., “Multilingual extremism detection in social media,” Expert Systems with Applications, 2021.
[21] Z. Waseem and D. Hovy, “Hateful conduct on Twitter: Annotated dataset,” in Proc. NAACL, 2016.
[22] S. Zannettouet al., “On the origins of memes and extremist propaganda,” in Proc. ICWSM, 2018.
[23] M. H. Ribeiro et al., “Evolving radicalization on YouTube,” in Proc. AAAI ICWSM, 2019.
[24] P. Neumann, “The trouble with extremism,” Perspectives on Terrorism, 2013.
[25] M. C. Benigni et al., “Online extremism and influence operations,” IEEE, 2017.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)