International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

Hybrid Lexicon-Driven News Threat Detection Using Random Forest and XGBoost Models

G Bhavani Prasad[1], G.Rajini[2]
[1]MCA Student , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad
[2]Assistant Professor , Dept of MCA , Ellenki College of Engineering and Technology , Hyderabad

***Abstract*— The rapid growth of online news platforms and social media has made it easier for information to spread quickly, but it has also increased the circulation of content that may create fear, misinformation, or potential security concerns. Detecting threat-related news at an early stage is essential for supporting public safety and informed decision-making. This study presents a hybrid lexicon-driven news threat detection framework that combines the NRC Emotion Lexicon with advanced machine learning models, namely Random Forest and XGBoost. Initially, news articles undergo preprocessing steps such as text cleaning, tokenization, and stop-word removal to improve data quality. Emotional features are then extracted using the NRC Lexicon and integrated with textual features to create an informative dataset for classification. The processed data is used to train and evaluate both machine learning models using performance measures including accuracy, precision, recall, and F1-score. Experimental results indicate that the proposed hybrid approach effectively identifies threat-related news, with the Random Forest model providing slightly better classification performance than XGBoost. The combination of emotion-based lexical analysis and ensemble learning enhances prediction accuracy, making the proposed framework a practical and reliable solution for intelligent news threat detection in real-world applications.**

***Keywords*— News Threat Detection, NRC Emotion Lexicon, Random Forest, XGBoost, Sentiment Analysis, Natural Language Processing (NLP), Feature Extraction, Machine Learning, Text Classification, Threat Intelligence.**

I. INTRODUCTION

The rapid expansion of digital news platforms and online communication has transformed the way people access and share information. Millions of news articles are published every day, allowing information to spread across the world within seconds. Although this has improved communication and public awareness, it has also increased the circulation of misleading, harmful, and threat-related content that can influence public opinion and social stability [2]. Threatening news articles may create fear, encourage violence, or spread misinformation, making early identification an important requirement for maintaining public safety and national security [1]. Manual monitoring of such a large amount of online content is both difficult and time-consuming. Therefore, intelligent text analysis techniques have become an important area of research for automatically identifying potentially harmful news articles. Recent advances in Natural Language Processing (NLP) and Machine Learning (ML) provide effective methods for analyzing textual information and supporting reliable threat detection in large-scale news environments [7].

Traditional text classification techniques generally depend on statistical representations such as Bag-of-Words and TF-IDF to analyze textual data. While these methods are effective in identifying frequently occurring words, they often fail to capture the emotional meaning associated with threat-related news [9]. News articles describing security incidents commonly contain emotions such as fear, anger, sadness, and anticipation, which provide valuable clues for identifying potential threats. Lexicon-based sentiment analysis has therefore gained considerable attention because it extracts emotional information directly from text and improves semantic understanding [11]. Researchers have also demonstrated that combining lexical knowledge with sentiment analysis techniques provides richer

feature representations than relying only on statistical text features [12]. These findings highlight the importance of integrating emotional information into machine learning models for more accurate threat classification.

To address these limitations, this study proposes a hybrid framework that combines the NRC Emotion Lexicon with TF-IDF features for news threat detection. Initially, news articles undergo preprocessing steps including text cleaning, tokenization, stop-word removal, punctuation elimination, and lemmatization to improve data quality. Emotion-related features such as fear, anger, joy, trust, sadness, surprise, disgust, and anticipation are then extracted using the NRC Emotion Lexicon, while TF-IDF generates numerical representations of important textual terms [15]. The integration of these two feature sets creates a comprehensive representation of both emotional and statistical characteristics within the news articles. Previous research has shown that combining lexicon-based methods with machine learning techniques significantly improves sentiment classification performance compared to individual approaches [8]. This hybrid representation provides a stronger foundation for identifying threat-related news accurately.

The generated hybrid feature vectors are used to train two advanced machine learning classifiers, namely Random Forest and XGBoost, to distinguish between threat-related and non-threat news articles. These models are selected because of their strong predictive performance, ability to process high-dimensional data, and robustness in handling complex classification problems. Their performance is evaluated using commonly accepted metrics such as accuracy, precision, recall, F1-score, confusion matrix, and classification report. Comparative evaluation enables the identification of the classifier that provides the most reliable threat prediction. Previous studies have demonstrated that machine learning techniques can significantly improve sentiment analysis and opinion mining performance when combined with appropriate feature engineering [16]. Similar research also highlights the effectiveness of integrating lexical information with classification algorithms to enhance prediction accuracy in text analysis applications [14].

The proposed framework provides a user-friendly environment where users can upload datasets, preprocess news articles, extract emotion-based features, train classification models, compare performance, and predict threats from newly entered news content. This automated workflow reduces manual effort while improving the speed and consistency of threat identification. Experimental evaluation demonstrates that the hybrid approach improves classification performance, with the Random Forest model achieving better overall results than XGBoost. The developed system offers a practical solution for supporting media monitoring, cybersecurity, public safety, and national security applications through early detection of potentially harmful news articles [19]. Furthermore, strengthening intelligent threat detection contributes to broader human security by assisting organizations in identifying risks before they spread widely across digital communication platforms [20].

II. RELATED WORK

Razali et al., [2021] [2] presented a detailed study on opinion mining for national security applications. Their work explored various sentiment analysis techniques and discussed how online textual data can be used to identify potential security threats. The authors compared lexicon-based and machine learning approaches and highlighted their advantages and limitations. They emphasized that combining multiple techniques improves threat detection accuracy and supports better decision-making. The study also discussed challenges such as handling large datasets and identifying complex emotional patterns. This research provides a strong foundation for developing intelligent systems that detect threat-related news using sentiment analysis and machine learning.

Ravi and Ravi, [2015] [7] reviewed different opinion mining and sentiment analysis methods used for text classification. Their study explained various preprocessing techniques, feature extraction methods, and machine learning algorithms applied to sentiment analysis. The authors showed that hybrid approaches combining lexical knowledge with machine learning produce better classification performance than individual methods. They also discussed applications of sentiment analysis in social media, business, and public opinion monitoring. Their work highlights the importance of effective feature selection and serves as a useful reference for developing accurate news threat detection systems.

Taj et al., [2019] [11] proposed a lexicon-based approach for sentiment analysis of news articles. Their study focused on identifying emotional information from textual content using predefined sentiment dictionaries. The authors showed that proper preprocessing and lexical feature extraction improve sentiment classification accuracy. They also explained that emotion-based analysis helps in understanding the actual meaning behind news articles. The experimental results demonstrated that lexicon-based techniques provide meaningful information for analyzing large collections of news

data. Their work supports the integration of emotion lexicons with machine learning models for intelligent news threat detection.

Vaitheeswaran and Arockiam, [2016] [15] introduced a hybrid sentiment analysis framework by combining lexicon-based methods with machine learning algorithms. Their research showed that integrating emotional features with statistical learning improves prediction accuracy. The authors emphasized that preprocessing and feature engineering play a major role in building efficient sentiment classification systems. Experimental results confirmed that the hybrid approach performed better than individual methods. This study demonstrates that combining lexical information with machine learning is an effective solution for analyzing large textual datasets and detecting meaningful patterns.

Zhang et al., [2014] [9] reviewed machine learning and lexicon-based approaches for sentiment classification. Their research compared the strengths of both techniques and discussed their applications in opinion mining. The authors found that machine learning models are effective in learning text patterns, while lexicon-based methods capture semantic and emotional information more effectively. They suggested that combining these approaches improves overall classification accuracy. Their study provides valuable guidance for designing hybrid text classification systems and serves as a strong reference for developing intelligent news threat detection frameworks.

III. DATASET DETAILS

The dataset used in this project contains news articles collected from various online sources and is designed to classify news content as either Threat or Non-Threat. Each record mainly includes the news text and its corresponding class label, which serves as the target variable for model training. The dataset covers different categories of news with diverse writing styles and topics, allowing the machine learning models to learn patterns associated with threatening and non-threatening content. It is organized in a structured tabular format, making it suitable for text processing and classification tasks. Before using the dataset, its contents are verified to ensure that the required fields are available and free from major inconsistencies. This dataset provides the foundation for extracting both textual and emotional information, enabling the proposed system to identify potential threats in news articles with improved accuracy.

Before training the models, several preprocessing techniques are applied to improve the quality of the textual data. The news articles are converted to lowercase, punctuation, special characters, and unnecessary words are removed, and the remaining text is tokenized and lemmatized to produce a standardized representation. The cleaned text is then transformed into TF-IDF features, while emotion-related features are extracted using the NRC Emotion Lexicon. These two feature sets are combined to create a hybrid feature representation that captures both textual patterns and emotional context. Finally, the dataset is divided into 80% training data and 20% testing data, allowing the Random Forest and XGBoost models to learn from historical data and evaluate their performance on unseen news articles. Proper dataset preparation ensures reliable and consistent threat detection results.

IV. PROPOSED METHODOLOGY

The proposed system follows a systematic workflow to detect threat-related news articles using a combination of Natural Language Processing (NLP), the NRC Emotion Lexicon, and advanced machine learning algorithms. Initially, the news dataset is uploaded into the system, where each record contains the news content and its corresponding label. The uploaded data is then preprocessed by converting text into lowercase, removing punctuation, special characters, numbers, and stop words, followed by tokenization and lemmatization to produce clean and standardized text. After preprocessing, emotion-based features such as anger, fear, sadness, joy, trust, disgust, surprise, and anticipation are extracted using the NRC Emotion Lexicon. These emotional features are combined with TF-IDF textual features to construct a hybrid feature representation that captures both the linguistic and emotional characteristics of news articles. The generated hybrid feature set is then prepared for machine learning model training and evaluation.

After feature construction, the hybrid dataset is divided into 80% training data and 20% testing data for model development and performance evaluation. Two supervised machine learning algorithms, namely Random Forest and XGBoost, are trained using the prepared dataset to classify news articles as Threat or Non-Threat. The performance of both models is assessed using evaluation metrics such as Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and Classification Report. A comparison graph is generated to visualize the performance of both classifiers and identify the better-performing model. Based on the experimental results, the Random Forest classifier achieves superior overall performance and is selected for final deployment. The trained model is then used to analyse newly entered news articles in real time, providing the predicted threat category along with a confidence

score to support accurate and reliable threat detection.

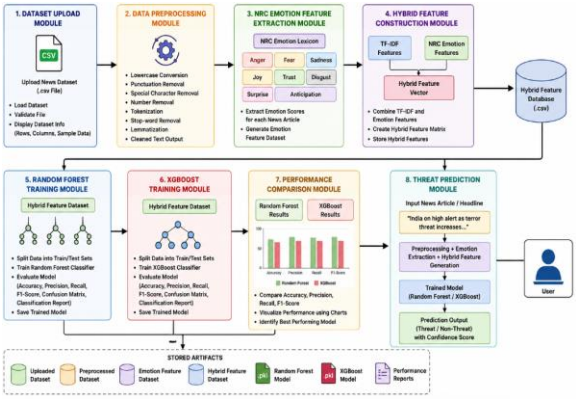

Figure [1]: News Threat Detection Framework

Figure [1] illustrates the workflow of the proposed News Threat Detection System. The process begins with dataset upload and text preprocessing, followed by NRC emotion feature extraction and hybrid feature construction using TF-IDF and emotion features. The hybrid dataset is used to train Random Forest and XGBoost models, whose performance is compared using standard evaluation metrics. Finally, the best-performing model predicts whether a new news article is Threat or Non-Threat along with its confidence score.

V. RESULT AND DISCUSSION

The experimental results demonstrate that the proposed News Threat Detection System can effectively identify threat-related news articles using a combination of text processing, emotion analysis, and machine learning techniques. The uploaded dataset is first cleaned through preprocessing, after which emotion scores are extracted using the NRC Emotion Lexicon. These emotional features are combined with TF-IDF features to create a hybrid feature set that improves the learning capability of the classification models. Random Forest and XGBoost are trained and tested using the prepared dataset, and their performance is evaluated with accuracy, precision, recall, and F1-score. The results indicate that both models perform well in classifying news articles, while the Random Forest model achieves better overall performance. The confusion matrices and classification reports show that most news articles are correctly classified with only a few prediction errors. The real-time prediction module further confirms the effectiveness of the trained model by accurately classifying new news articles as Threat or Non-Threat.

```
Random Forest training completed!
Accuracy  : 0.9722
Precision : 0.9737
Recall    : 0.9722
F1 Score  : 0.9722
```

Figure [2]: Random Forest Model Results

Figure [2] presents the evaluation results of the Random Forest classifier using the hybrid feature dataset. The system displays the confusion matrix and classification report, showing that the model correctly classifies most news articles with only a few prediction errors. These results indicate that the Random Forest model is effective for news threat detection.

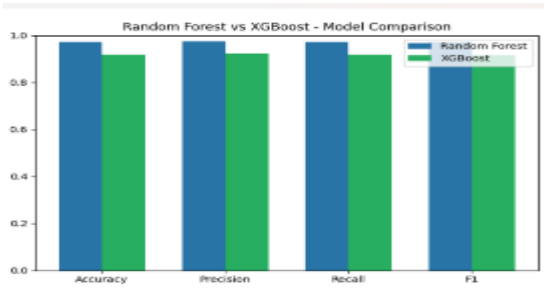

Figure [3]: Performance Comparison Graph

Figure [3] illustrates the performance comparison between the Random Forest and XGBoost classifiers using evaluation metrics such as accuracy, precision, recall, and F1-score. The graph provides a clear visual comparison of both models and shows that the Random Forest classifier performs better overall, making it more suitable for news threat detection.

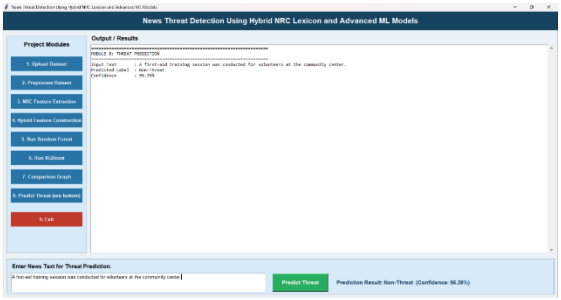

Figure [4]: Threat Prediction Output

Figure [4] shows the final prediction generated by the proposed News Threat Detection System. After the user enters a news article, the system preprocesses the text, extracts the required features, and classifies it as either Threat or Non-Threat. The prediction result is displayed along with a confidence score, enabling users to quickly understand the classification outcome.

DISCUSSION

The outcome of this project shows that combining textual features with emotion-based information improves the accuracy of news threat detection. Text preprocessing removes unwanted information from news articles, allowing the models to learn more meaningful patterns from the data. The inclusion of NRC emotion features helps capture emotional expressions that are commonly associated with threat-related content, making the classification process more effective. Among the evaluated models, Random Forest provides better results than XGBoost across all performance measures, indicating its suitability for this application. The comparison graph and confusion matrix clearly illustrate the difference in model performance and support the selection of Random Forest as the final classifier. The system also performs well in real-time prediction by analysing new news articles and providing the predicted category with a confidence score. Overall, the proposed framework offers an efficient and dependable approach for identifying potentially threatening news content in online media.

VI. CONCLUSION

The proposed News Threat Detection Using Hybrid NRC Lexicon and Advanced Machine Learning Models provides an effective approach for identifying whether a news article is Threat or Non-Threat. The system combines text preprocessing, TF-IDF feature extraction, and NRC Emotion Lexicon-based emotion analysis to generate a hybrid feature set that captures both textual and emotional characteristics of news content. These features are used to train and evaluate the Random Forest and XGBoost classifiers, allowing a detailed comparison of their performance. The experimental results show that both models perform well, while the Random Forest classifier delivers more consistent and reliable predictions. The developed application offers a simple graphical interface through which users can upload datasets, preprocess data, train models, compare their performance, and predict the threat level of new news articles. The prediction module also provides a confidence score, helping users better understand the classification results. Overall, the proposed framework offers an accurate and practical solution for automated news threat detection and can support organizations involved in media monitoring, cybersecurity, and public safety by enabling faster identification of potentially harmful news content.

REFERENCES

[1] J. R. Clapper, "Statement for the Record: Worldwide Threat Assessment of the U.S. Intelligence Community," Office of the Director of National Intelligence, Congressional Testimonies, USA, 2015. [Online]. Available: https://www.dni.gov/files/SFRDirNCTCSHSGAC Hearing8Oct.pdf

[2] N. A. M. Razali et al., "Opinion Mining for National Security: Techniques, Domain Applications, Challenges and Research Opportunities," Journal of Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00536-5.

[3] S. Dorle, "Sentiment Analysis Methods and Approaches: A Survey," International Journal of Innovative Computer Science and Engineering, vol. 4, no. 6, pp. 1–5, Dec. 2017. [Online]. Available: http://www.ijicse.in/index.php/ijicse/article/view/134

[4] A. Balahur, R. Steinberger, E. Van Der Goot, B. Pouliquen, and M. Kabadjov, "Opinion Mining on Newspaper Quotations," in Proc. IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Sep. 2009, pp. 523–526, doi: 10.1109/WIIAT.2009.340.

[5] B. Seerat, "Opinion Mining: Issues and Challenges—A Survey," International Journal of Computer Applications, vol. 49, no. 9, pp. 42–51, 2012, doi: 10.5120/7658-0762.

[6] P. Barnaghi, J. G. Breslin, I. D. A. B. Park, and L. Dangan, "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment," in Proc. IEEE 2nd International Conference on Big Data Computing Service and Applications (BigDataService), Mar.–Apr. 2016, pp. 52–57.

[7] K. Ravi and V. Ravi, "A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications," Knowledge-Based Systems, vol. 89, pp. 14–46, Nov. 2015.

[8] G. Isabelle, W. Maharani, and I. Asror, "Analysis on Opinion Mining Using Combining Lexicon-Based Method and Multinomial Naïve Bayes," in Proc. International Conference on Industrial Enterprise Systems Engineering, vol. 2, 2019, pp. 214–219, doi: 10.2991/icoiese-18.2019.38.

[9] H. Zhang, W. Gan, and B. Jiang, "Machine Learning and Lexicon-Based Methods for Sentiment Classification: A Survey," in Proc. 11th Web Information Systems and Applications Conference (WISA), Sep. 2014, pp. 262–265, doi: 10.1109/WISA.2014.55.

[10] M. P. Ashna and A. K. Sunny, "A Study on Sentiment Analysis in Malayalam Language," International Journal of Advanced Research in

Computer and Communication Engineering, vol. 6, Special Issue 3, pp. 88–93, 2017.

[11] S. Taj, B. B. Shaikh, and A. Fatemah Meghji, "Sentiment Analysis of News Articles: A Lexicon-Based Approach," in Proc. 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Jan. 2019, doi: 10.1109/ICOMET.2019.8673428.

[12] X. Ding, B. Liu, and P. S. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining," in Proc. International Conference on Web Search and Web Data Mining (WSDM), 2008, pp. 231–239, doi: 10.1145/1341531.1341561.

[13] B. Saberi and S. Saad, "Sentiment Analysis or Opinion Mining: A Review," International Journal of Advanced Science, Engineering and Information Technology, vol. 7, no. 5, pp. 1660–1666, 2017, doi: 10.18517/ijaseit.7.5.2137.

[14] W. A. Zgheib and A. M. Barbar, "A Study Using Support Vector Machines to Classify the Sentiments of Tweets," International Journal of Computer Applications, vol. 170, no. 2, pp. 8–12, Jul. 2017, doi: 10.5120/IJCA2017914690.

[15] G. Vaitheeswaran and D. L. Arockiam, "Combining Lexicon and Machine Learning Methods to Enhance the Accuracy of Sentiment Analysis on Big Data," International Journal of Computer Science and Information Technologies, vol. 7, no. 1, pp. 306–311, 2016.

[16] A. D'Andrea, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, Tools and Applications for Sentiment Analysis Implementation," International Journal of Computer Applications, vol. 125, no. 3, pp. 26–33, Sep. 2015, doi: 10.5120/ijca2015905866.

[17] S. L. Cosby, "Human Security Concept: The Root of U.S. National Security and Foreign Policy," United States Marine Corps, USA, 2009.
[18] S. Stoyko and G. S. Rakovski, "Management, Security and the System of National Security," Military Academy, Sofia, Bulgaria, 2011.

[19] T. Balzacq, "What Is National Security?" Revue Internationale et Stratégique, vol. 52, no. 4, pp. 33–50, 2015, doi: 10.3917/ris.052.0033.

[20] B. Bahadur and S. Thakuri, "Human Security: Concept, Dimensions and Challenges," Institute of Crisis Management Studies, Kathmandu, Nepal, 2018.

[21] Shashank, A. (2025). AI-Enhanced ETL Processes: Leveraging Artificial Intelligence for Optimized Data Integration Systems. Journal Of Multidisciplinary, 5(8), 219-225.