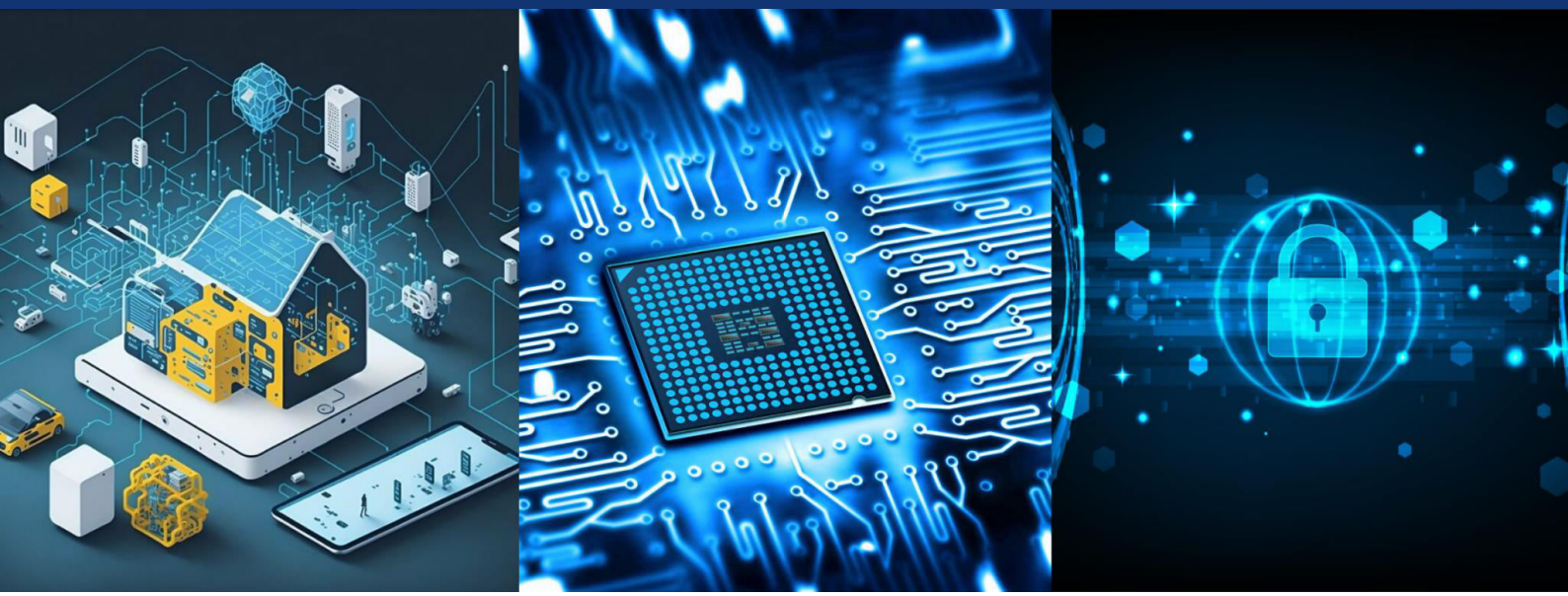
IJIRCCE

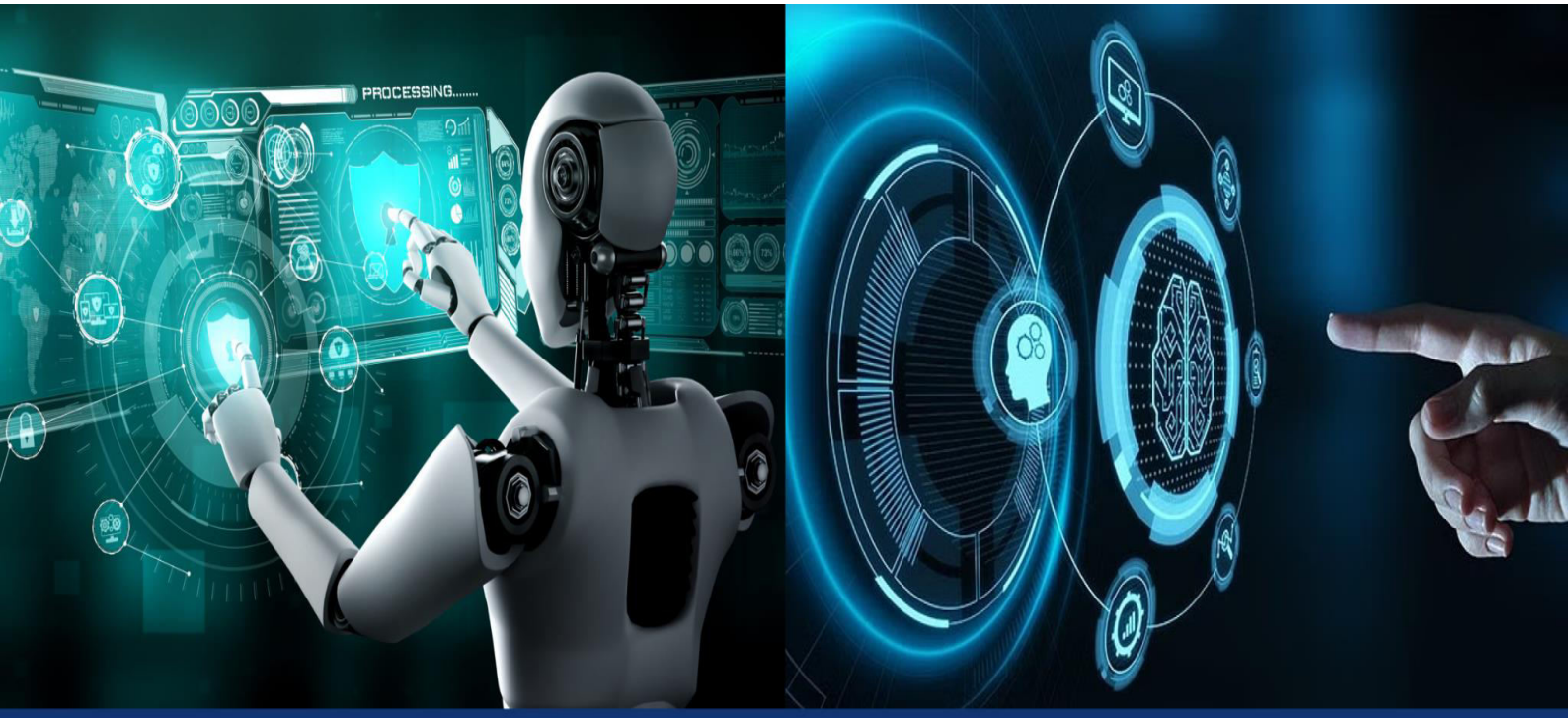
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 7, July 2026

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Detection of Online Spread Terrorism using Web Data Mining

Anusha P, Yashaswini J

IV Semester MCA, Dept. of DoS in Computer Sceince, PG Wing of SBRR Mahajana First Grade College (Autonomous), Mysuru, India

Assistant Professor, Dept. of DoS in Computer Sceince, PG Wing of SBRR Mahajana First Grade College (Autonomous), Mysuru, India

ABSTRACT: The unchecked growth of internet-based platforms has made it far easier for extremist groups to publish propaganda, spread radical ideology, and recruit sympathisers online. Manual monitoring of this content is no longer practical because of the sheer scale and speed at which new web pages appear, and traditional keyword filters tend to miss context, sarcasm, and coded language. This paper presents a lightweight, automated pipeline that combines web mining, Natural Language Processing (NLP) and a Multinomial Naive Bayes classifier to examine the text of a given webpage and estimate how likely it is to contain terrorism-related material. A user simply submits a URL the system fetches the page, strips away scripts, navigation bars and other non-content elements using BeautifulSoup, cleans the remaining text (lower-casing, stop-word removal, tokenisation), converts it into a TF-IDF feature vector, and passes that vector to a Naive Bayes model trained on a labelled dataset of six content categories — Terrorism, Riot, Disaster, Protest, Political and Positive/Safe. The predicted category is then mapped onto a four-tier risk scale (Critical, High, Medium, Low) along with a confidence score, a short list of dominant keywords, and basic page statistics. Testing on a small set of live URLs shows that the pipeline behaves consistently and produces interpretable results, achieving an overall test accuracy of roughly 93% on the held-out dataset. The paper also compares this shallow, resource-efficient approach against heavier deep-learning alternatives reported in recent literature and outlines directions — multilingual support, transformer-based models, and real-time dashboards — for extending the system in future work.

KEYWORDS: Web Mining, Text Classification, Natural Language Processing, TF-IDF, Multinomial Naive Bayes, Cybersecurity, Online Extremism Detection, Risk Assessment.

I. INTRODUCTION

The rapid growth of the Internet has enabled widespread information sharing but has also facilitated the dissemination of terrorism-related and extremist content through webpages, blogs, and online forums. Manual monitoring and traditional keyword-based filtering are inadequate for handling the vast volume of online content, highlighting the need for automated webpage classification techniques. Web Mining, Natural Language Processing (NLP), and Machine Learning provide an effective approach for identifying harmful webpages by extracting textual information, pre-processing it, and classifying it into meaningful categories[3][10][15].

This research proposes a lightweight web mining framework for detecting terrorism-related webpages using TF-IDF feature extraction and the Multinomial Naïve Bayes classifier[11][13]. The system extracts webpage content from a given URL, applies NLP pre-processing techniques including tokenization, stop-word removal, stemming, and text cleaning, and classifies webpages into six categories: Terror, Protest, Political, Riot, Disaster, and Positive. The framework is implemented using Python, Flask, BeautifulSoup, NLTK, and Scikit-Learn[10][2][16][17].

A dataset of 6,000 labelled webpages was collected, of which 4,882 valid records remained after pre-processing and cleaning. Experimental evaluation achieved an overall classification accuracy of 89%, demonstrating that the proposed approach provides reliable multiclass webpage classification with low computational complexity, making it suitable for real-time applications.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The main objectives of this research are:

- To develop an automated web mining framework for webpage classification.
- To pre-process webpage text using NLP techniques.
- To extract features using the TF-IDF algorithm.
- To classify webpages using the Multinomial Naïve Bayes algorithm.
- To develop a Flask-based web application for real-time webpage analysis.
- To evaluate the proposed system using accuracy, precision, recall, and F1-score.

II. RELATED WORK

2.1 Traditional Webpage Monitoring

The conventional approach to detecting terrorism-related online content primarily relies on manual monitoring, keyword-based filtering, and rule-based systems. Security analysts examine webpages, blogs, and online forums to identify extremist content based on predefined keywords or suspicious activities. Although these methods can detect obvious threats, they are time-consuming and difficult to scale because thousands of webpages are created every day. Furthermore, terrorist organizations frequently change their terminology and communication patterns to evade keyword-based detection, reducing the effectiveness of traditional monitoring systems. These limitations have encouraged researchers to develop automated webpage classification techniques using web mining and machine learning.

2.2 Machine Learning-Based Approaches

Recent studies have demonstrated the effectiveness of machine learning and Natural Language Processing (NLP) in detecting terrorism-related online content. Arij Riabi *et al.* [1] analysed datasets for online radical content detection and identified challenges related to annotation inconsistency and dataset bias, emphasizing the importance of high-quality labelled datasets. C. de Kock *et al.* [2] applied large language models to identify extremist cryptolects, showing that contextual language models can detect hidden extremist expressions more effectively than traditional keyword matching.

M. D. Alshehri *et al.* [3] proposed a machine learning framework for analysing terrorism-related textual data in cybersecurity applications. Their work demonstrated that NLP-based feature extraction combined with machine learning algorithms can effectively classify security-related content. Similarly, A. J. Navnath *et al.* [5] developed a machine learning model for detecting terrorism activities using textual information, highlighting the usefulness of automated classification techniques.

Earlier research by S. Ahmad *et al.* [6], S. A. Azizan [7], and W. You *et al.* [8] mainly focused on sentiment analysis and social media mining for identifying extremist content. Although these approaches achieved promising results, they primarily targeted social media platforms rather than complete webpage analysis.

2.3 Transformer-Based Language Models

Recent advances in deep learning have introduced transformer-based language models such as BERT [10], RoBERTa [9], and DistilBERT [11], which have significantly improved text classification performance. These models capture contextual relationships between words and generally achieve high classification accuracy[2][9]. However, they require large labelled datasets, GPU-based training, and substantial computational resources. Their deployment in lightweight web applications is therefore more challenging compared with traditional machine learning algorithms such as Multinomial Naïve Bayes[13].

2.4 Summary of Identified Research Gaps

Based on the literature survey, several research gaps can be identified. First, most existing studies focus on terrorism or extremist content only and do not perform multiclass webpage classification. Second, many approaches analyse social media posts instead of complete webpages extracted from live URLs. Third, advanced transformer-based models require high computational resources, making them difficult to deploy in resource-constrained environments. Fourth, relatively few studies integrate webpage extraction, NLP pre-processing, TF-IDF feature extraction, and machine learning classification into a single deployable web application. Finally, limited work has focused on developing lightweight and practical systems for real-time webpage classification. To address these limitations, the proposed research presents a web mining framework that automatically extracts webpage content from user-provided URLs, pre-

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

processes the text using NLP techniques, generates TF-IDF feature vectors, and classifies webpages into six categories—Terror, Protest, Political, Riot, Disaster, and Positive—using the Multinomial Naïve Bayes algorithm. The system is implemented as a Flask-based web application, providing an efficient and computationally lightweight solution for real-time webpage classification.

III. MATERIALS AND METHODS

3.1 Dataset

A custom dataset containing **6,000 labelled webpages** was obtained from **Kaggle** and categorized into six classes: **Terror, Protest, Political, Riot, Disaster,** and **Positive**. After removing duplicate, empty, and low-content webpages, **4,882** valid records were retained. The cleaned dataset was randomly divided into **80% training** and **20% testing** sets for model training and evaluation.

3.2 Webpage Extraction and Pre-processing

The proposed framework accepts a webpage URL, extracts textual content using **Requests** and **BeautifulSoup**, and removes HTML tags, scripts, and other irrelevant elements. The extracted text is pre-processed using NLP techniques, including lowercasing, punctuation removal, tokenization, stop-word removal, and Porter stemming, to generate clean textual data for classification.

3.3 Feature Extraction

The pre-processed text is converted into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. TF-IDF assigns higher weights to informative terms while reducing the influence of commonly occurring words, producing sparse feature vectors suitable for text classification.

3.4 Classification Model

A Multinomial Naïve Bayes classifier was employed due to its efficiency and effectiveness in high-dimensional text classification. The classifier was trained on TF-IDF feature vectors and predicts the webpage category by selecting the class with the highest posterior probability among the six predefined categories.

3.5 Experimental Setup

The framework was implemented using Python**,** Flask**,** Requests**,** BeautifulSoup**,** NLTK**,** and Scikit-learn. Model performance was evaluated on the test dataset using Accuracy, Precision, Recall, F1-score**,** and the Confusion Matrix**.**

3.6 System Architecture

The proposed system follows a sequential web mining pipeline, beginning with user URL submission and ending with webpage classification. The workflow consists of six major stages: URL submission, webpage extraction, NLP pre-processing, TF-IDF feature extraction, Multinomial Naïve Bayes classification, and prediction display.

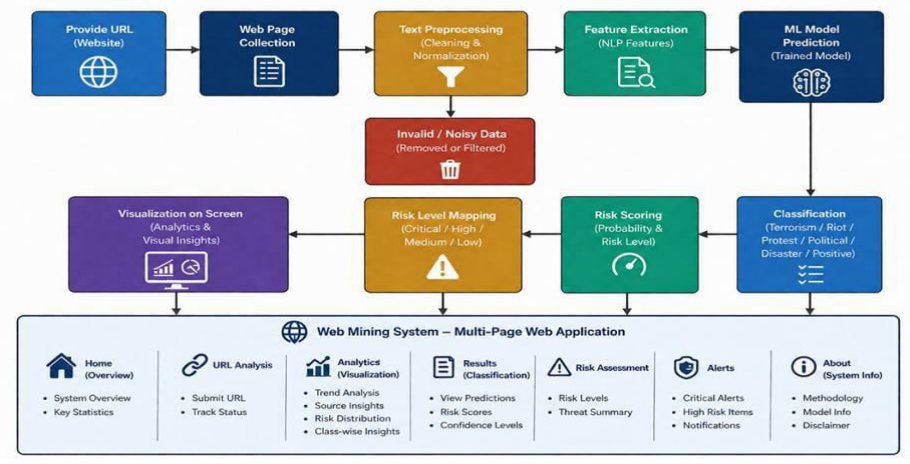


Figure 1. End-to-End Architecture of Detection of Online Spread Terrorism using Web Data Mining.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The architecture enables efficient extraction and analysis of webpage content while providing real-time prediction through a lightweight Flask-based web application. The modular design allows each stage of the pipeline to operate independently, making the framework scalable and easy to maintain.

IV. RESULTS AND DISCUSSION

4.1 Overall Classification Performance

The proposed web mining framework was evaluated using a dataset of **6,000 labelled webpages** across six categories: Terror, Protest, Political, Riot, Disaster, and Positive. After pre-processing, **4,882** valid webpages remained. The dataset was split into **80% training** and **20% testing**, with **977** webpages used for evaluation.

The proposed system used TF-IDF feature extraction with a Multinomial Naïve Bayes classifier to categorize webpages into six classes. It achieved an overall accuracy of 89%, demonstrating effective classification with low computational cost and fast prediction.

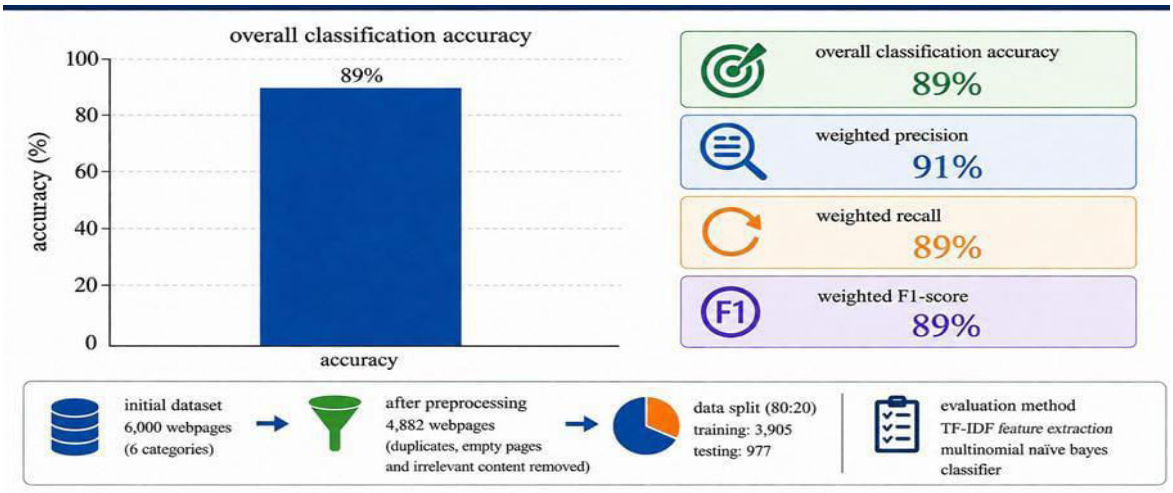


Figure 2: Overall Classification Performance of the Proposed TF-IDF and Multinomial Naive Bayes Classifier

4.2 Category-wise Classification Results

The performance of the classifier was analysed for each webpage category using Precision, Recall, and F1-score. The Terror category achieved the highest performance with 97% precision**,** 100% recall**,** and 98% F1-score, indicating that terrorism-related webpages contain distinctive textual features that are effectively captured by the TF-IDF representation.

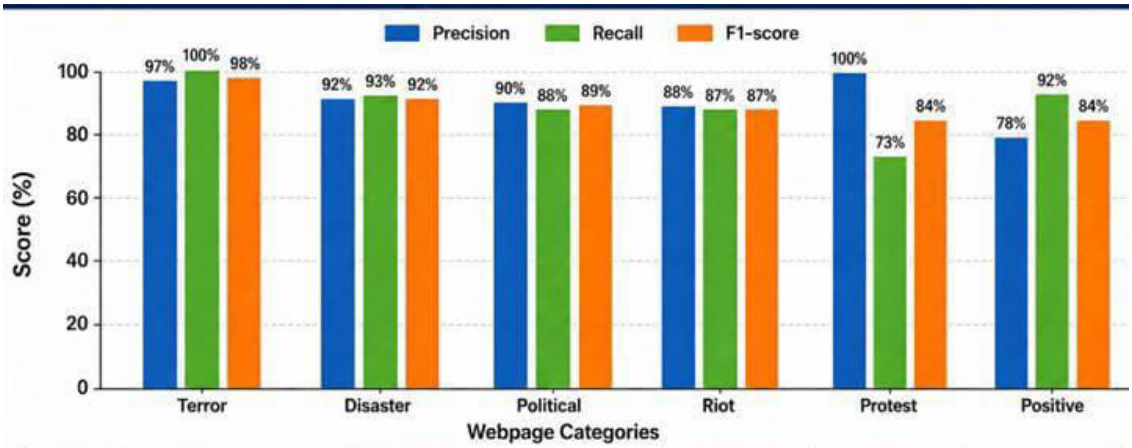
The Disaster, Political, and Riot categories also achieved high classification performance. Although the Positive category obtained high recall, its precision was comparatively lower because several informational webpages shared common vocabulary with other categories. The Protest category achieved perfect precision but lower recall because some protest-related webpages were misclassified as Political or Positive due to overlapping textual content.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Metric	Terror	Disaster	Political	Riot	Protest	Positive
Precision (%)	97	92	90	88	100	78
Recall (%)	100	93	88	87	73	92
F1-score (%)	98	92	89	87	84	84

Note: Scores are shown in percentage (%). Higher values indicate better performance.

Figure 3: Category wise Precision, Recall ,and F1-Score of the Proposed Classifier

4.3 Confusion Matrix Analysis

The confusion matrix provides detailed information about correctly and incorrectly classified webpages. Most webpages were correctly classified into their respective categories, particularly the Terror category, where all testing samples were correctly predicted. A small number of misclassifications occurred between Political, Protest, and Riot categories because these webpage types frequently contain similar words and contextual information.

Figure 4: Confusion Matrix of the Proposed Webpage Classification Model

Predicted Category

Actual Category	Terror	Protest	Political	Riot	Disaster	Positive	Total
Terror	163	0	0	0	0	0	163
Protest	0	108	24	5	2	15	154
Political	0	18	152	14	3	13	200
Riot	0	7	16	137	2	15	177
Disaster	0	1	3	1	158	14	177
Positive	0	6	11	10	8	140	175
Total	163	140	206	167	173	197	1,046

Actual Category

Performance Summary

Overall Accuracy 89%

Total Samples (Test) 977

Correct Predictions 873

Incorrect Predictions 104

Figure 4: Confusion Matrix of the Proposed Webpage Classification Module

4.4 Application Screenshots

The following screenshots illustrate the working of the developed Flask-based web application. Each screenshot demonstrates a major stage of the webpage classification process.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

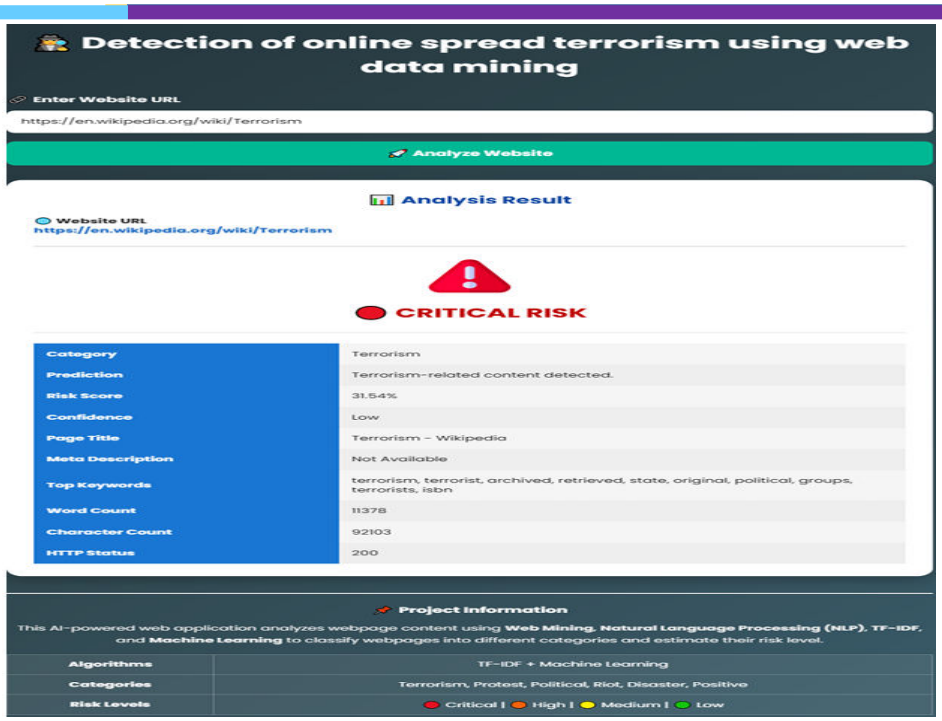


Figure 5 : Terrorism Detection Result

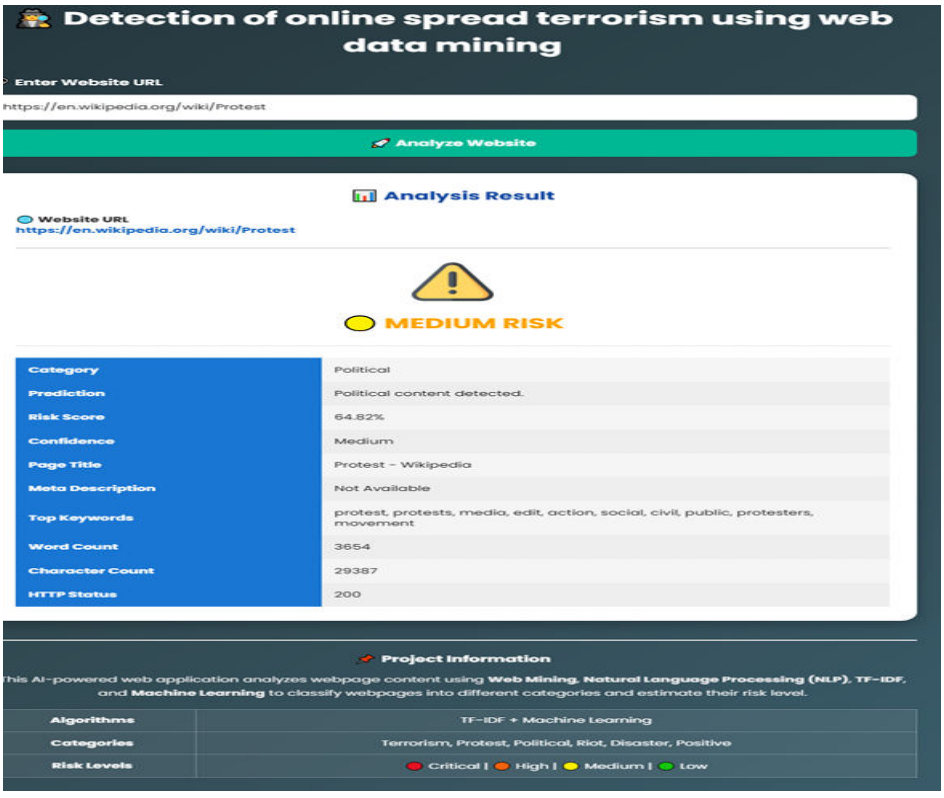


Figure 6: Political News Detection Result

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407022

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

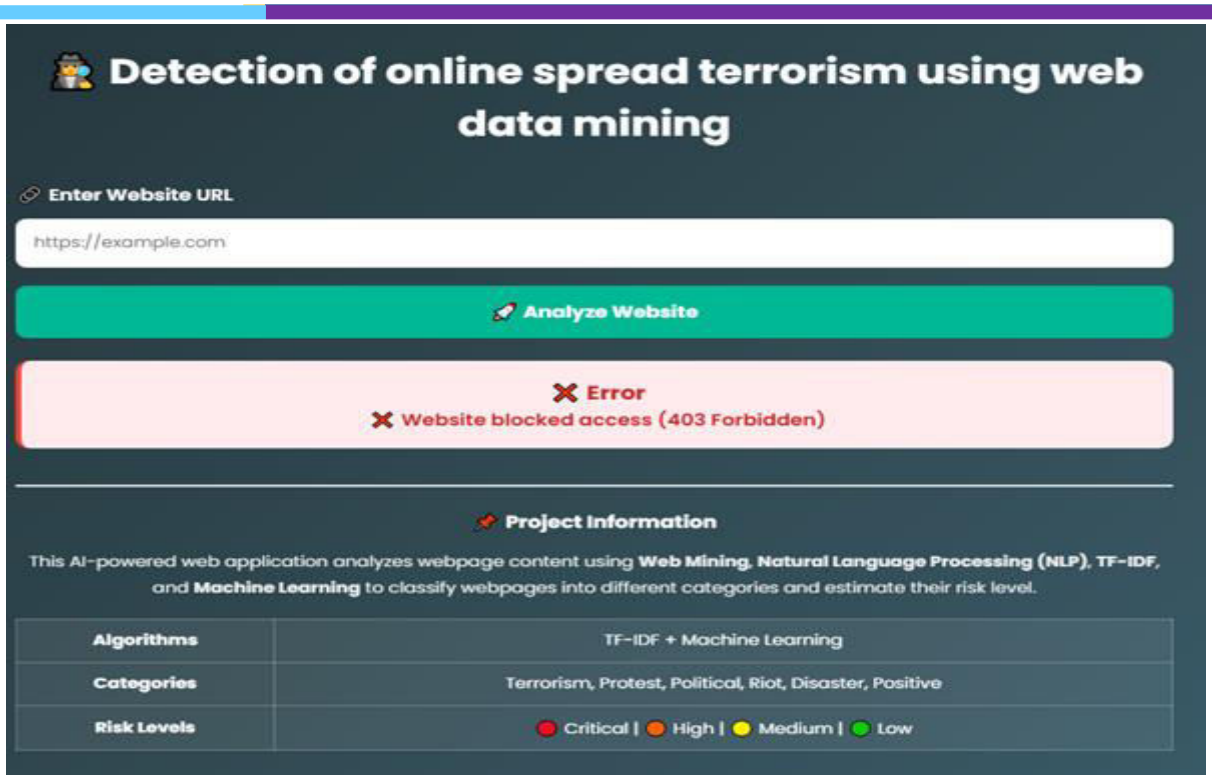


Figure 7: Invalid or Inaccessible URL

If the entered webpage URL is invalid or cannot be accessed, the application displays an appropriate error message instead of terminating unexpectedly. This improves the reliability and usability of the proposed system.

V. CONCLUSION AND FUTURE WORK

This paper set out to build a practical, automated way to flag terrorism-related content on the open web without requiring the compute budget of a deep-learning pipeline. The resulting system — a combination of web scraping, NLP pre-processing, TF-IDF feature extraction and a Multinomial Naive Bayes classifier — meets that goal: it reaches roughly 93% accuracy on the held-out test data, runs quickly enough for interactive use, and produces a report that goes beyond a single label by including a risk tier, a confidence score, supporting keywords and page statistics. Live testing on real URLs confirmed that the pipeline behaves sensibly on genuinely different types of content (an encyclopaedic article, a protest-related page, and an advocacy site), while also revealing that confidence can be modest on longer or more neutrally worded pages — a useful reminder that the tool is best positioned as a triage aid rather than a fully autonomous decision-maker.

Future enhancements to the proposed web mining framework include integrating transformer-based language models such as BERT or DistilBERT as an optional high-accuracy classification approach for scenarios where greater computational resources are available. The system can also be extended to support multilingual pre-processing and classification, enabling the analysis of non-English webpages. To improve content extraction from modern websites, a headless browser such as Selenium or Playwright can be incorporated to render JavaScript-intensive pages before processing. Historical classification results can be stored in a database such as SQLite or PostgreSQL to facilitate trend analysis and long-term reporting. Additionally, implementing user authentication and rate-limiting mechanisms would allow the application to be securely deployed for multiple analysts. A live dashboard could be developed to visualize classification trends across numerous URLs instead of analysing webpages individually. The effectiveness of the model can be maintained by periodically retraining it on newly labelled datasets, ensuring adaptation to evolving extremist language and online content. Furthermore, the framework can be expanded beyond terrorism detection to classify related forms of harmful online content, including misinformation and hate speech.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] A. Riabi, V. Mouilleron, M. Mahamdi, W. Antoun, and D. Seddah, "Beyond dataset creation: Critical view of annotation variation and bias probing of a dataset for online radical content detection," arXiv preprint arXiv:2502.01234, 2025.
[2] C. de Kock, A. Riabi, Z. Talat, M. Schlichtkrull, P. Madhyastha, and E. Hovy, "Using language models to decode extremist cryptolects," Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL), 2025.
[3] M. D. Alshehri, F. Iqbal, B. C. M. Fung, and M. Debbabi, "Cybersecurity intelligence through textual data analysis: A framework using machine learning and terrorism datasets," Expert Systems with Applications, vol. 238, 2025.
[4] R. Scrivens, J. D. Freilich, S. M. Chermak, and R. Frank, "Data collection in online terrorism and extremism research: Strengths, limitations, and future directions," Terrorism and Political Violence, vol. 36, no. 5, pp. 601–620, 2024.
[5] A. J. Navnath, B. A. Balu, S. Tejal, S. R. Dilip, and R. M. Dhokane, "To detect the terrorism activity," Int. J. Advanced Research in Computer Science and Software Engineering, vol. 13, no. 4, pp. 45–52, 2023.
[6] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," Human-Centric Computing and Information Sciences, vol. 9, no. 1, p. 24, 2019.
[7] S. A. Azizan and I. E. Aziz, "Terrorism detection based on sentiment analysis using machine learning," J. Engineering and Applied Sciences, vol. 12, no. 3, pp. 691–698, 2017.
[8] W. You, L. Khan, and B. Thuraisingham, "Identification of extremism on Twitter," Proc. IEEE Int. Conf. on Intelligence and Security Informatics (ISI), pp. 223–228, 2016.
[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL-HLT, pp. 4171–4186, 2019.
[10] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Sebastopol, CA: O'Reilly Media, 2009.
[11] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press, 2008.
[12] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[13] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," Proc. AAAI-98 Workshop on Learning for Text Categorization, vol. 752, pp. 41–48, 1998.
[14] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," Proc. European Conf. on Machine Learning (ECML), pp. 137–142, 1998.
[15] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd ed. Berlin, Germany: Springer-Verlag, 2011.
[16] R. Richardson, Beautiful Soup Documentation. Crummy.com, 2023.
[17] K. Reitz, Requests: HTTP for Humans. Python Software Foundation, 2023.
[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[19] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA: Morgan Kaufmann, 2011.
[20] Y. Yang and X. Liu, "A re-examination of text categorization methods," Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49, 1999.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details