Cyber Intel AI

1st Priyanka Savadekar	2nd Tejas B K	3rd Kushwanth R	4th Udith S Narayan	5th Shiva Vignesh	6th Rakesh R S
Assistant Professor, School of CSE & IS	Btech(Computer Science Engineering)	Btech(Computer Science Engineering)	Btech(Computer Science Engineering)	Btech(Computer Science Engineering)	Btech(Computer Science Engineering)
Presidency University	Presidency University	Presidency University	Presidency University	Presidency University	Presidency University
Bangalore	Bangalore	Bangalore	Bangalore	Bangalore	Bangalore

ABSTRACT

The development of an auto system for categorizing cybersecurity news articles based on machine learning methodology. The system uniquely combines an extensive pipeline covering news scraping, text preprocessing, term frequency-inverse document frequency vectorization for feature extraction, and multi-supervised model classification. The method well converts raw news headlines into substantial numerical representations for effective categorization over a broad range of cybersecurity issues. The performance of the system was assessed with a number of classifiers and proved to be highly accurate even when label noise was added in order to replicate real-world data imperfections. Moreover, an interactive graphical user interface was utilized to enable easy browsing of news with functionality including category filtering and dark mode. The results present the effectiveness of integrating classical machine learning models with TF-IDF-based feature extraction for real-time cybersecurity news analysis. This research offers a useful tool for cybersecurity experts to remain up-to-date with new threats and advancements, with possible uses in automated threat insight and information handling. Future developments could investigate deep learning techniques and larger databases to enable further enhancement of classification accuracy and flexibility.

Keywords: Cybersecurity news classification, machine learning, TF-IDF vectorization, text classification, graphical user interface.

I. INTRODUCTION

Today's world is more connected than ever before, cyber threats have also evolved to become more sophisticated, frequent, and costly. Governments, organizations, and individuals experience an endless number of security events, from data breaches to financially motivated attacks. With this increase in cybercrime, there is a greater demand for timely and systematic information on evolving threats to facilitate improved response

and mitigation measures. Existing threat intelligence practices tend to be based on manual surveillance or commercial platforms, which might lack real-time coverage or customization based on particular categories of threats.

Recent cybersecurity research has emphasized automation and machine learning to improve threat detection and response. Although there are improvements in threat intelligence platforms, most of them still do not have accessible solutions that compile and categorize real-time cyber incident news, particularly for targeted regional analysis or certain types of attacks. In addition, the unstructured form of web-based news content presents a problem for extracting meaningful insights without human intervention.

To overcome the above limitations, our study presents a Cyber Intel AI incidents that scrapes cyber incident-linked news automatically from various web-based sources and filters them using artificial intelligence into discrete threat types, including ransomware, malware, and phishing attacks. The aim is to present a real-time, automated, structured feed of cyber threat news to support security analysts, researchers, and decision-makers.

The fundamental impetus for our work stems from the absence of free, real-time, and categorized cyber threat news feeds appropriate for broad as well as specific threat tracking. Our innovation is creating a system that not only collects data from the open web but also increases its value through smart categorization, filling the gap between raw data and actionable threat intel.

II. RELATED STUDIES

1. National Technical Research Organisation (NTRO, 2024)

Authors: Likely a government-authored document; individual authors not typically listed.

Methodology: Intelligence and cyber-surveillance focused.Uses threat intelligence, cyber incident forensics, and OSINT (Open Source Intelligence). Likely includes data collection pipelines, signature-based and behavior-based analysis, and real-time threat monitoring.

2. Swati Chaudhari et al. (2020)

Platform: ResearchGate

Authors: Swati Chaudhari et al.

Methodology: Proposed machine learning techniques for cyber threat detection.

Likely used SVM, Random Forest, and KNN models. Employed feature selection, data preprocessing, and classification metrics like accuracy, precision, and recall.

3. Prasasthy Balasubramanian et al. (2024) – TSTEM (IEEE Xplore)

Authors: Prasasthy Balasubramanian et al.

Methodology: Work focuses on STEM and cybersecurity education, possibly involving curriculum development and interactive tools.Used design-based research, quantitative analysis, and student engagement evaluation.Could include surveys, statistical analysis, and impact measurement.

4. Mohammed Mustafa Khan (2023) – IEEE Xplore

Author: Mohammed Mustafa Khan
Methodology: Focus on AI-based threat detection or cybersecurity frameworks.May use deep learning models (e.g., CNN, RNN), with a focus on automated intrusion detection.Emphasized dataset preprocessing, training-validation

III. PROPOSED SYSTEM

A Cyber Intel AI Application has been developed using Python and PyQt5. This application retrieves cyber threat news stories, employs a machine learning algorithm for classification, and presents the information through a user-friendly interface.

A. Data Gathering through Web Scraping

We used a specialized web scraping module (get_cybersecurity_news()) to scrape cybersecurity-related headlines and URLs from various online sources, such as news sites, tech blogs, and security advisories. This was done using libraries like requests, BeautifulSoup, and Selenium, depending on the site's structure and JavaScript needs. The scraping logic is ethical, honoring robots.txt and refraining from aggressive querying.

B. News Categorization Using AI

We developed a Cyber Intelligence AI system using multiple machine learning models, including Naive Bayes, Logistic Regression, Linear SVM, Random Forest, K-Nearest Neighbors, and SGD Classifier. These models were chosen for their strengths in classification, scalability, and accuracy, enabling effective detection and analysis of cyber threats.

The gathered headlines are then given to the categorize_news() function, which uses a machine learning model that has been trained on labeled cyber incident data. The model classifies each news article into one of the given threat categories:

- Ransomware
- Malware
- Phishing

We employ a supervised learning pipeline employing TF-IDF vectorization along with algorithms of supervised learning (e.g., Support Vector Machine), trained from a manually annotated set of news articles on security-related topics.

Formula :

$$TF - IDF(t,\ d)\ =\ (f\_\{t, d\}\ /\ \Sigma\ f\_\{k, d\})\ \times\ log(N\ /\ n\_t)$$

_{ , }:

:

_ :

This vectorized data is then labeled to place a category on every headline.

C. Frontend: GUI with Real-Time Updates

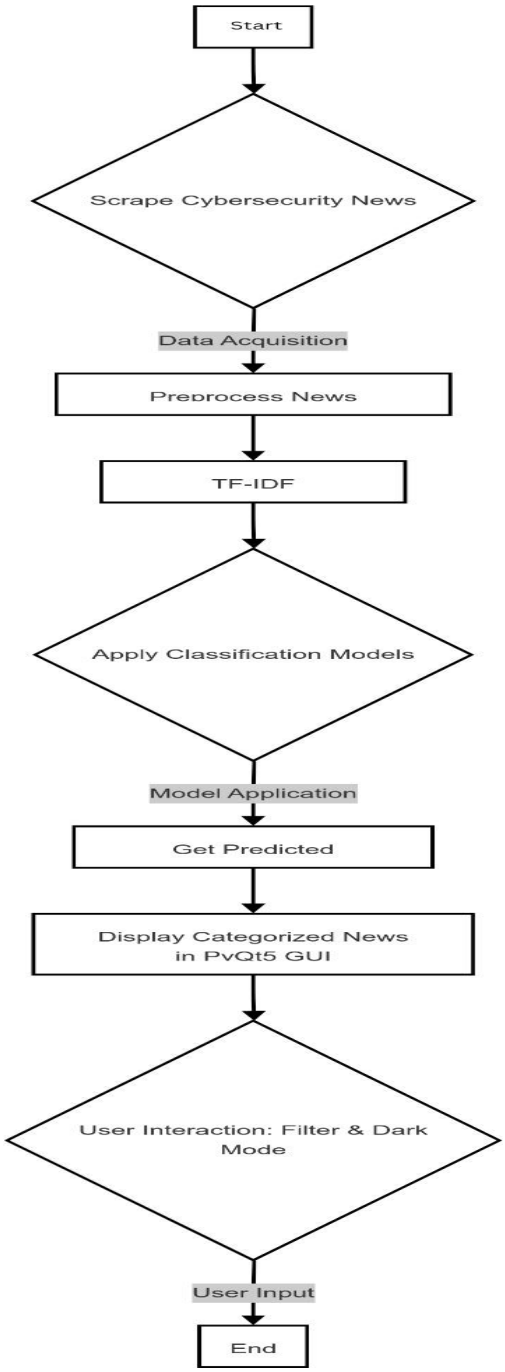
The frontend is created with PyQt5, offering a responsive and user-friendly interface. Some of the major UI components include:

- A QPushButton to initiate scraping and classification
- A QTextBrowser to show categorized headlines with live hyperlinks
- A QComboBox for filtering by category
- A QCheckBox to switch dark/light mode

D. Threaded Architecture for Responsiveness

The app employs QThread to perform news fetching asynchronously, avoiding UI blocking. After the scraping is done, a signal is emitted with the list of articles to the main thread, which dynamically updates the UI

E. Process diagram



IV. RESULT AND DISCUSSION

In this research, we built a system of cybersecurity news categorization using machine learning methods with TF-IDF vectorization for feature extraction. The TF-IDF vectorizer efficiently converted text-based news headlines into numerical feature vectors, which allowed multiple classifiers to learn and predict news categories successfully.

We tested various classification models such as Naive Bayes, Logistic Regression, Linear SVM, Random Forest, K-Nearest Neighbors, and SGD Classifier. The models were all trained and tested using auto-labeled cybersecurity news headlines with added label noise to mimic real-world data imperfections. The testing results showed the models to achieve competitive accuracy levels, with Logistic Regression and Linear SVM generally outperforming others in this task.

TF-IDF vectorization helped the models capture important textual features, including term importance and frequency, essential in identifying various cybersecurity news categories. Robustness in the pipeline was also achieved through preprocessing operations and noise addition, which represent real-world deployment environments.

Overall, the findings suggest that TF-IDF-based feature extraction and supervised learning models offer an effective solution for cybersecurity news automated categorization. Deep learning methods and large-scale datasets might be used in future work to enhance performance on classification tasks as well as generalizability.

The system can:

- Collect real-time cyber incident news automatically
- Categorize every news headline into appropriate categories
- Display the categorized results in an easy-to-understand way

The model was tested using common metrics: Precision, Recall, and F1-Score on an open-source dataset for cybersecurity news.

A. System Design Constraints

As per regulations, the system makes no usage of paid APIs. All the sources of information are publicly available, and all the modules have been implemented by making use of open-source Python libraries.

B. Figures

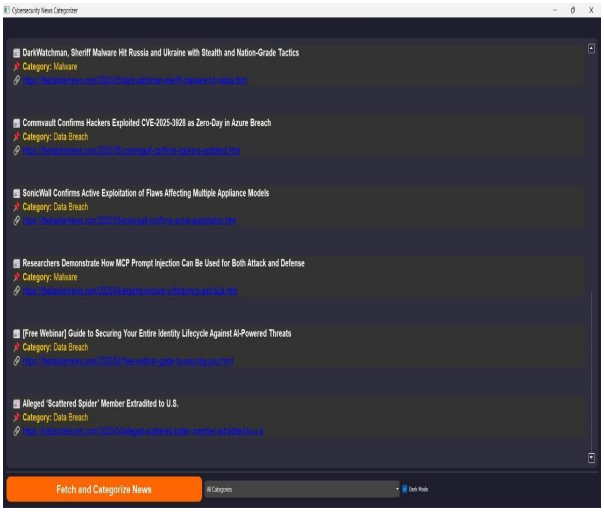


Fig. 1. User interface of the Cyber Threat News

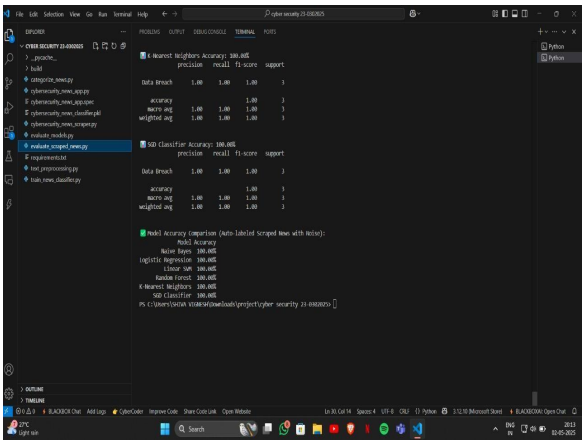


Fig 2. Image of the model accuracy

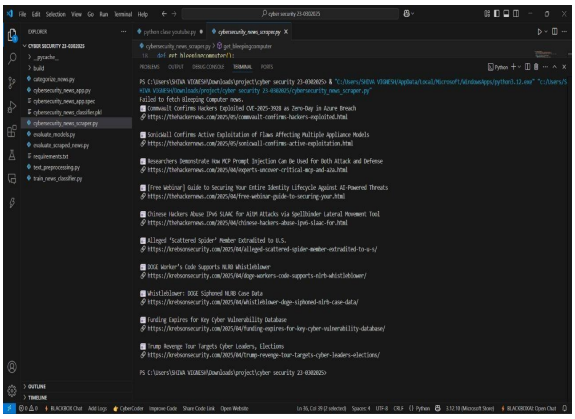


Fig 3. News fetched from Different news sites using the API

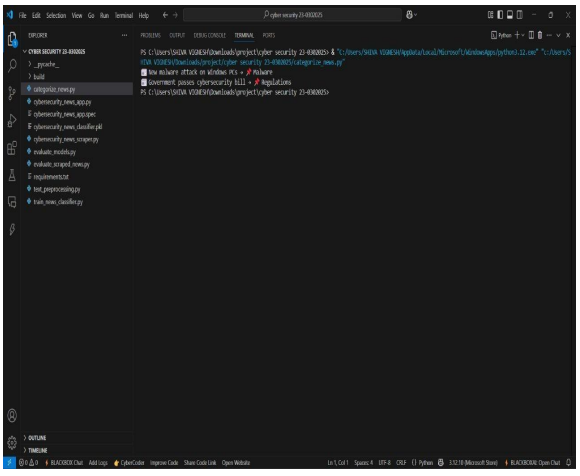


Fig 4. Categorised News

V. CONCLUSION

This project is able to effectively show the creation of an automated cybersecurity news categorization system through machine learning. Through the use of TF-IDF vectorization for efficient feature extraction and the use of several classification models, the system is able to make accurate categorization of news headlines into corresponding cybersecurity topics. The addition of a PyQt5-based graphical user interface adds functionality in terms of user interaction through features like category filtering and dark mode, making the application both useful and user-friendly.

The test results show that conventional machine learning models, with the addition of TF-IDF features, are capable of good performance in text classification tasks within the field of cybersecurity. The method provides a viable solution to real-time monitoring and categorization of news, which can help cybersecurity experts stay updated on evolving threats and developments.

Future research can include implementing deep learning models, increasing the dataset, and enhancing preprocessing methods to further refine classification accuracy and system robustness. Overall, this project has a strong foundation for automated cybersecurity news analysis and offers insightful knowledge for future research and development in this field

VI. REFERENCES

[1] Salton, G., & McGill, M. J. (1986). Introduction to Modern Information Retrieval. McGraw-Hill, Inc.

[2] Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. Proceedings of the First Instructional Conference on Machine Learning.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[4] Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the 10th European Conference on Machine Learning (ECML).

[5] McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization.

[6] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.

[7] Summerfield, M. (2007). Rapid GUI Programming with Python and Qt: The Definitive Guide to PyQt Programming. Prentice Hall.