A REPORT
ON

Cyber Intel AI

Submitted by,

Rakesh R S	**– 20211CCS0152**
Kushwanth R	**– 20211CCS0147**
Udith S Narayan	**– 20211CCS0164**
Tejas B K	**– 20211CCS0183**
Chintalapudi Shiva Vignesh	**– 20211CCS0178**

Under the guidance of,

Ms. Priyanka Niranjan Savadekar

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

At

PRESIDENCY UNIVERSITY

BENGALURU

MAY 2025

PRESIDENCY UNIVERSITY

PRESIDENCY SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Internship/Project report **"Cyber Intel AI"** being submitted by "Rakesh R S, Kushwanth R, Udith S Narayan, Tejas B K, Chintalapudi Shiva Vignesh" bearing roll numbers "20211CCS0152, 20211CCS0147, 20211CCS0164, 20211CCS0183, 20211CCS0178" in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering(Cyber Security) is a Bonafide work carried out under my supervision.

Ms. Priyanka Savadekar
Assistant Professor
School of CSE&IS
Presidency University

Dr. S P Anandaraj
Professor & HOD
School of CSE&IS
Presidency University

Dr. MYDHILI NAIR
Associate Dean
PSCS
Presidency University

Dr. SAMEERUDDIN KHAN
Pro-Vice Chancellor - Engineering
Dean –PSCS / PSIS
Presidency University

PRESIDENCY UNIVERSITY

PRESIDENCY SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

I hereby declare that the work, which is being presented in the report entitled "**Cyber Intel AI"** in partial fulfillment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering**, is a record of my own investigations carried under the guidance of **Ms. Priyanka Niranjan Savadekar, Assistant Professor, Prsidency School of Computer Science and Engineering, Presidency University, Bengaluru. .**

I have not submitted the matter presented in this report anywhere for the award of any other Degree.

Names	Roll No	Signature
Rakesh R S	20211CCS0152	
Kushwanth R	20211CCS0147	
Udith S Narayan	20211CCS0164	
Tejas B K	20211CCS0183	
Chintalapudi Shiva Vignesh	20211CCS0178	

PROJECT COMPLETION CERTIFICATE

- **The certificate issued from an organization must have the duration of the Internship, i.e.start and end date, project title and a technology on which work is carried out.**

ASTRACT

In today's increasingly digital world, cyber threats are evolving at an unprecedented rate, posing significant risks to individuals, organizations, and nations. Timely access to accurate and relevant information about cybersecurity incidents is critical for maintaining robust cyber defense mechanisms. Our project, titled "Cyber Intel AI," addresses this need by developing a comprehensive system that aggregates, classifies, and presents real-time cyber incident news from multiple credible online sources.

The system comprises several core components, beginning with a web scraping module that continuously extracts news articles from leading cybersecurity websites. The collected data is then passed through a preprocessing pipeline, where it is cleaned, structured, and prepared for analysis. To enhance the relevance and quality of the information presented to users, we incorporate a machine learning-based classification model. This model categorizes news articles based on malware types (e.g., ransomware, spyware, trojans) and filters out irrelevant or duplicate content using natural language processing techniques.

A notable feature of our platform is the malware-specific filtering engine, which empowers users to easily navigate the vast amount of cybersecurity news by selecting categories most relevant to their interests or operational needs. The user interface has been designed to be intuitive, visually clean, and responsive, providing users with a seamless experience as they browse real-time cyber incident updates.

Overall, the tool not only saves time and effort by automating the collection and classification of incident news but also contributes to improved situational awareness and threat intelligence. By delivering targeted, up-to-date, and categorized cyber incident feeds, this system serves as a valuable resource for cybersecurity professionals, analysts, and organizations seeking to stay ahead of emerging threats.

ACKNOWLEDGEMENTS

First of all, we are indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Pro-VC - Engineering and Dean, Presiency School of Computer Science and Engineering & Presiency School of Information Science, Presidency University for getting us permission to undergo the project.

We express our heartfelt gratitude to our beloved Associate Dean **Dr. Mydhili Nair,** Presidency School of Computer Science and Engineering, Presidency University, and "**Dr. Anandaraj S P**", Head of the Department, Presidency School of Computer Science and Engineering, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide **Ms. Priyanka Niranjan Savadekar Assistant Professor** and Reviewer **Dr./Mr. Mohana S D, Assistant Professor**, Presidency School of Computer Science and Engineering, Presidency University for his/her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the internship work.

We would like to convey our gratitude and heartfelt thanks to the CSE7301 Internship/University Project Coordinator **Mr. Md Ziaur Rahman and Dr. Sampath A K,** department Project Coordinators **DR.Sharmasth Vali,** and Git hub coordinator **Mr. Muthuraj.**

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

RAKESH R S(1)

KUSHWANTH R(2)

UDITH S NARAYAN(3)

TEJAS B K(4)

CHINTALAPUDI SHIVA VIGNESH(5)

LIST OF FIGURES

Sl. No.	Figure Name	Caption	Page No.
1.	Figure 1	Flow chart----------	**10**
2.	Figure 2	Architecture diagram----------	**11**
3.	Figure 3	Gantt chart----------	**13**
4.	Figure 4	Categorize news----------	**25**
5.	Figure 5	News Scraper----------	**25**
6.	Figure 6.1	Evaluate Scraped News----------	**26**
	Figure 6.2	Evaluate Scraped News----------	**26**
	Figure 6.3	Evaluate Scraped News----------	**27**
	Figure 6.4	Evaluate Scraped News----------	**27**
7.	Figure 7.1	Evaluate Model----------	**28**
	Figure 7.2	Evaluate Model----------	**28**
8.	Figure 8	User interface----------	**29**
9.	Figure 9	News app----------	**30**
10.	Figure 10	Sustainable Development Goals mapping---------	**34**

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	Abstract	I
	Acknowledgment	II
1.	Introduction	1
	1.1 Background	1
	1.2 Motivation	1
	1.3 Project Overview	1
2.	Literature survey	2
3.	Research gaps of existing methods	6
4.	Proposed methodology	7
	4.1 Web Scraping	7
	4.2 Data Preprocessing	7
	4.3 Machine Learning Classification	7
	4.4 Filtering Engine	8
	4.5 Interface	8
5.	Objectives	9
6.	System design & implementation	10
	6.1 Web Scraper Module	12
	6.2 Auto Labelling Module	12
	6.3 Data Preparation Module	12
	6.4 Machine Learning Module	12
	6.5 Evaluation and Reporting Module	12
7.	Timeline for execution of project (Gantt chart)	13
8.	Outcomes	15
9.	Results and discussions	16
10.	Conclusion	18
	References	19
	Appendix A, B, C	22-35

Chapter 1

INTRODUCTION

1.1 Background

1.1.1 Evolution of Cyber Threats

In the modern digital era, cyber threats have become increasingly sophisticated, frequent, and impactful. With organizations and individuals heavily reliant on digital infrastructure, the importance of real-time cyber threat intelligence cannot be overstated. Timely awareness and response to these threats are essential in mitigating potential damage and safeguarding sensitive data.

1.2 Motivation

The manual process of gathering threat intelligence from various websites is inefficient and prone to oversight. There is a growing need for a system that can automatically collect, process, and display relevant cybersecurity news in real-time. This project aims to fill that gap through an automated tool that offers malware-based filtering and machine learning-enhanced news relevance classification.

1.3 Project Overview

The project titled "Cyber Intel AI" is a web-based platform that aggregates cybersecurity incident news from multiple trusted sources. The tool allows users to filter news based on different malware categories and leverages a machine learning model to classify and prioritize relevant information. The platform provides a user-friendly interface for cybersecurity professionals to stay informed about emerging threats.

Chapter 2

LITERATURE SURVEY

Author(s) & Year	Title	Methodology	Strengths	Limitations
National Technical Research Organisation (NTRO) 2024	Developing a Tool for Real-Time Cyber Incident Feeds for Indian Cyber Space	• Gathered cyber incident data through web crawling and scraping from social media and forums. • Applied machine learning to categorize incidents and created a real-time monitoring dashboard.	• India-Specific Focus • Real-Time Monitoring • Automation and Intelligence	• Data Quality and Noise • Scalability and Performance • Dependence on Internet Accessibility
Swati Chaudhari, 2020	Real-Time Logs and Traffic Monitoring, Analysis and Visualization Setup for IT Security Enhancement	• Employed the ELK stack to gather and analyze actual log and network traffic information. • Focused on identifying serious security events to avoid data breaches and compromise..	Effective use of the ELK stack for real-time analysis and visualization of security-related data; focus on preventing major information security disasters.	The study does not provide specific accuracy metrics; implementation details specific to Indian cyberspace are limited.
Prasasthy Balasubraman (2024)	TSTEM: A Cognitive Platform for Collecting Cyber Threat Intelligence in the Wild	Developed a containerized microservice architecture using tools like Tweepy, Scrapy, Terraform, ELK, Kafka, and MLOps; employed custom crawlers and NLP models for classification and entity extraction to autonomously search, extract, and index Indicators of Compromise (IOCs).	High accuracy in classification and extraction tasks; efficient multi-stage IOC extraction methodology; real-time processing capabilities.	While the platform demonstrates high accuracy, its application is not specifically tailored to Indian cyberspace; potential challenges in handling diverse data sources are not detailed.
Mohammed Mustafa Khan 2023	Proactive Cyber Defense: Conducting Real-Time Monitoring and Analysis of Security Events Using SIEM Tools	Discusses the implementation of Security Information and Event Management (SIEM) tools for real-time monitoring and analysis of security events to detect and respond to potential security incidents.	Provides a holistic view of an organization's security landscape by aggregating data from various sources.	Does not provide specific accuracy metrics or detailed dataset information.
Aviral Srivastava,2023	Anticipated Network Surveillance: An Extrapolated	Proposes a novel technique to predict upcoming attacks in a network based on several data parameters using machine	Enhances network security by predicting potential attacks,	Specific accuracy metrics are not provided; implementation may

	Study to Predict Cyber-Attacks Using Machine Learning and Data Analytics	learning and data mining techniques. The model comprises dataset pre-processing, training, and testing phases.	allowing for proactive defense measures.	require significant computational resources.
Yohan Fernandes, 2024	Analysing India's Cyber Warfare Readiness and Developing a Defence Strategy	Explores network security behaviors through simulation models; implements a cyber threat detection system using machine learning to identify and categorize cyber threats in real-time; proposes integration strategies into India's existing infrastructure.	Provides a comprehensive analysis of India's cyber defense readiness and proposes both technological and educational solutions to enhance real-time threat detection and response capabilities.	Specific accuracy metrics are not provided; implementation may require significant changes to existing infrastructure.
Precious Gold, 2025	AI-Driven Threat Intelligence: Enhancing SIEM Capabilities for Real-Time Cybersecurity Monitoring	The paper explores how AI-powered threat intelligence improves SIEM systems by identifying zero-day vulnerabilities, reducing false positives, and enabling automated incident response mechanisms.	Highlights the integration of AI to enhance real-time threat detection and response in SIEM systems.	Discusses challenges such as computational power requirements and balancing automation with human intervention.
E Kocyigit, M Korkmaz, 2020	Real-Time Content-Based Cyber Threat Detection with Machine Learning	Proposes a cyber threat detection system that examines web page content using machine learning to classify pages as legitimate or malicious.	Emphasizes the dynamic nature of machine learning approaches in detecting malicious web content.	Specific limitations are not detailed in the summary.
Hamed Taherdoost 2024	Insights into Cybercrime Detection and Response: A Review of Time	Examines various approaches and quantitative measurements to understand the relationship between detection and response times in cybersecurity.	Provides a comprehensive review of existing literature, offering insights into effective strategies for cybercrime detection and response.	As a review paper, it may not present original experimental data
Prasasthy Balasubraman 2024	A Cognitive Platform for Collecting Cyber Threat Intelligence in the Wild	Implements a containerized microservice architecture using tools like Tweepy, Scrapy, Terraform, ELK, Kafka, and MLOps. Utilizes advanced Natural Language Processing models for	The platform provides efficient and accurate real-time detection, collection, and sharing of cyber	The system requires significant computational resources and may face challenges in processing vast amounts of data from

		classification and entity extraction to enhance IOC extraction accuracy.	threat intelligence, enhancing the resilience of IT and OT environments against large-scale cyber-attacks.	diverse sources.
Emre Koçyiğit, 2021	Developing a Tool for Real-Time Cyber Incident Feeds for Indian Cyber Space	The proposed system analyzes the content of web pages using machine learning techniques to determine their legitimacy. The approach involves feature extraction from web page content and training classifiers to detect malicious pages in real-time.	The content-based analysis allows for the detection of threats that traditional URL-based methods might miss, enhancing the system's robustness against evolving phishing techniques.	The lack of detailed information about the dataset and feature extraction process may affect the reproducibility and generalizability of the results.
Yohan Fernandes, 2024	AI-Driven Threat Intelligence: Enhancing SIEM Capabilities for Real-Time Cybersecurity Monitoring	Focuses on implementing a cyber threat detection system that uses machine learning to identify and categorize cyber threats in real-time. Proposes strategies to integrate this system into India's existing infrastructure and suggests an educational framework for training cyber professionals.	Provides a comprehensive analysis of India's cyber defense readiness and offers practical solutions, including technological and educational strategies, to enhance cybersecurity.	The study may face challenges in real-world implementation due to existing infrastructure limitations and the need for extensive training programs
Manikanta Pradeep Adupa, 2021	Implementation of Cyber Talk With Real-Time Insights Hub Using Machine Learning and Psutil App Framework	The study presents "CyberTalk with Real-time Insights Hub," a platform that integrates machine learning models with real-time threat intelligence. The system analyzes incoming data to provide predictive insights into emerging cyber threats, enabling proactive defense measures.	The platform's real-time analysis and predictive capabilities enable organizations to anticipate and mitigate potential threats before they materialize.	The absence of detailed accuracy metrics and dataset information limits the ability to assess the system's effectiveness fully.

With the evolving nature of cyber threats at a fast pace, several research studies have been conducted based on the growing need for timely threat detection and response mechanisms. A study conducted by the National Technical Research Organisation (NTRO, 2024) rightly highlights the creation of tools for real-time cyber incident feeds related to India cyberspace, in order to enable cyber awareness and defence at the national level. As per Swati Chaudhari et al. (2020), the system proposed identifies and displays logs and network traffic in real time as per the goal of improving internal IT security. TSTEM developed by Prasasthy Balasubramanian et al. (2024) provides an intellectual model for gathering cyber threat intelligence from changing and open web sources, whereas Mohammed Mustafa Khan (2023) depicts employment of SIEM tools for proactive real-time surveillance of security incidents. Furthermore, Aviral Srivastava et al. have documented cyberattack predictions using machine learning and data analysis that came on board in 2023. The authors also elaborated on how one may develop a tool providing real-time cyber feeds with focus on content-based filtering by Y. Fernandes and N. Abosata (2021). Following these advancements, Yohan Fernandes and Nasr Abosata incorporated artificial intelligence into the conventional SIEM tools to provide real-time threat detection.

Manikanta Adupa et al. (2021) give a definition of the machine learning enabled real-time monitoring hub in the Psutil platform for real-time insights into system health and system operation. Conversely, Hamed Taherdoost (2024) gives a detailed account of the majority of cybercrime detection and attribution techniques and hence filling the theory-to-practice gap. Similarly, Precious Gold et al. (2025) also describe the significant use of AI in threat categorization and prioritization. Fernandes and Abosata also concentrate on India's readiness in cyber warfare and suggest AI-facilitated upgradation to national defense. Certain of the entries, including those by Srivastava et al. and Fernandes et al., are duplicate entries, contributing to again emphasize the adoption of predictive analytics and AI toward fortifying the cybersecurity environment. Combined, these reports signal a shift toward proactive, intelligent, and adaptive cyber defense systems that facilitate real-time situational awareness and national cyber resilience.

Chapter 3
RESEARCH GAPS OF EXISTING METHODS

Existing malware classification and threat intelligence tools often suffer from several significant limitations that impact their effectiveness and user adoption. One major issue is the limited scope of malware classification, which is typically confined to basic or predefined categories. This restricts the system's ability to accurately identify and classify emerging or complex threats. Furthermore, many of these tools lack real-time update capabilities, resulting in a noticeable delay in the availability of crucial threat information, which can be detrimental in rapidly evolving cybersecurity landscapes. Another key shortcoming is the minimal integration of machine learning technologies to automatically filter and assess the relevance of incoming data, leading to information overload and inefficient threat prioritization. Additionally, these tools often provide inadequate filtering and customization options, making it difficult for users to tailor the platform to their specific needs or focus areas. The user experience is frequently compromised by non-intuitive interfaces and cumbersome workflows, which hinder productivity and deter consistent usage. Moreover, there is a notable lack of support for data visualization and trend analysis, limiting users' ability to interpret patterns, monitor evolving threats, and make informed decisions based on historical and real-time data. These gaps highlight the urgent need for more intelligent, adaptive, and user-centric threat intelligence solutions**.**

Chapter 4

PROPOSED METHODOLOGY

The proposed system uses a modular approach comprising the following steps:

4.1 Web Scraping:

An automated web scraping framework is used in this study to gather new cybersecurity news articles from several reliable internet sources. The scraper is made to effectively extract the most recent headlines, guaranteeing a current and varied dataset. To capture a broad spectrum of cybersecurity themes and viewpoints, multiple sources are explored. In order to preserve consistency throughout the dataset, the scraping process include managing pagination, eliminating unnecessary content, and normalizing the collected text.

4.2 Data Preprocessing:

A pre-trained classification algorithm that assigns topic labels based on patterns acquired from previously labeled data is used to auto-label the gathered headlines. Label noise is added by randomly changing a predetermined percentage of the labels in order to more accurately replicate real-world situations where labeling errors and inconsistencies arise. This stage aids in assessing how resilient classification models are to faulty data. To maintain the original class distribution, which is essential for objective model evaluation, the dataset is then divided into training and testing subsets using stratified sampling.

4.3 Machine Learning Classification:

Six machine learning classifiers are chosen from a broad range to see how well they classify cybersecurity news. These include optimization-based Stochastic Gradient Descent (SGD) Classifier, instance-based learning with K-Nearest Neighbors (KNN), ensemble techniques like Random Forest, probabilistic models like Naive Bayes, and linear models like Logistic Regression and Linear Support Vector Machine (SVM). By recording term frequency and inverse document frequency data, each model is integrated into a pipeline that uses TF-IDF vectorization to convert the textual headlines into numerical feature vectors. The models can identify discriminative patterns in the text data thanks to this representation.

4.4 Filtering Engine:

Each model is trained on the preprocessed training data and then tested on the unseen test set as part of the filter engine's training and evaluation pipeline. To give a thorough evaluation of each classifier's prediction skills, performance metrics such as accuracy, precision, recall, and F1-score are calculated. To evaluate performance by class and pinpoint strengths and shortcomings, comprehensive classification reports are produced. To make it easier to choose the best model for cybersecurity news classification tasks, the findings are compiled in a comparative table that takes into account both accuracy and robustness to label noise.

4.5 Interface:

The assessment pipeline and the user can interact more easily thanks to the system interface. Raw cybersecurity news headlines can be entered into the interface manually or automatically through scraping. The input is then processed by the noise injection and auto-labeling modules before being fed into the machine learning models. Users may quickly evaluate the efficacy of the model because to the interface's simple and understandable outputs, which include categorization reports, accuracy ratings, and comparison performance tables. Furthermore, the interface's modular design makes it simple to integrate other parts or future modifications, such adding more classifiers, integrating new data sources, or implementing the top-performing model in a real-time news filtering application.

Chapter 5

OBJECTIVES

The project's main goal is to create an automated system that can gather cybersecurity news in real time and make sure consumers have access to the most up-to-date and pertinent information. Using machine learning algorithms to categorize news articles into malware-specific groups is a crucial part of this system, which improves the relevance and specificity of the content that is supplied. In order to increase usability and user engagement. The dashboard of the system would be dynamic and adaptable, enabling users to customize their experience to suit their own requirements and tastes. In order to guarantee that the system consistently detects significant risks while reducing noise, a great focus will be on attaining high precision and recall in the filtering of pertinent news articles. The ultimate goals of this technology are to raise user awareness, facilitate quicker and better-informed incident response within cybersecurity operations, and expedite access to vital threat intelligence.

The purpose of the "evaluate_scraped_news.py" script is to evaluate how well different machine learning models classify cybersecurity news items. A pre-trained categorization algorithm is used to automatically categorize the headlines after it has first scraped recent cybersecurity news from various sources. The script optionally adds label noise by randomly changing some of the labels to mimic faults in real-world data.After that, the data is divided into testing and training sets. Using a TF-IDF vectorizer to translate text into numerical features, a number of classification models are defined and trained on the training data, including Naive Bayes, Logistic Regression, Linear SVM, Random Forest, K-Nearest Neighbors, and SGD Classifier. Accuracy scores and thorough categorization reports are produced once each model's performance is assessed on the test set. The script concludes by displaying a summary table that contrasts each model's accuracy and offers information on how well each one performs in classifying cybersecurity news.

Chapter 6

SYSTEM DESIGN & IMPLEMENTATION

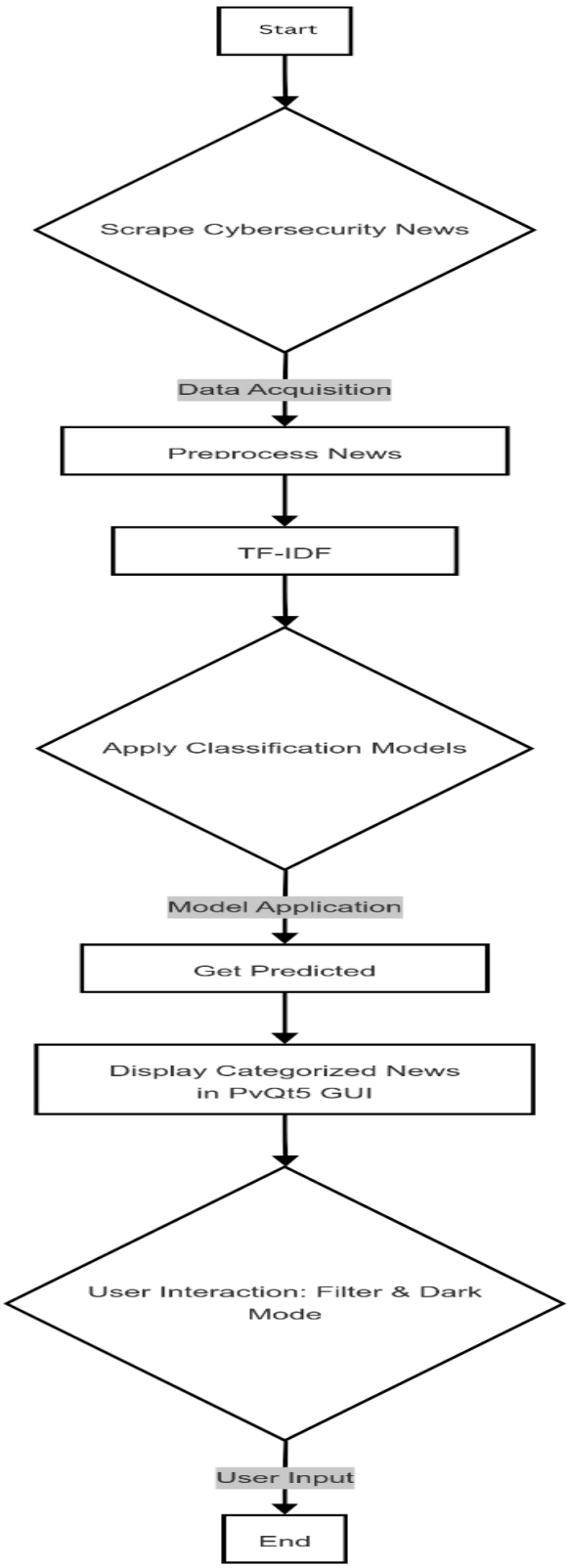


Fig1. Flow chart

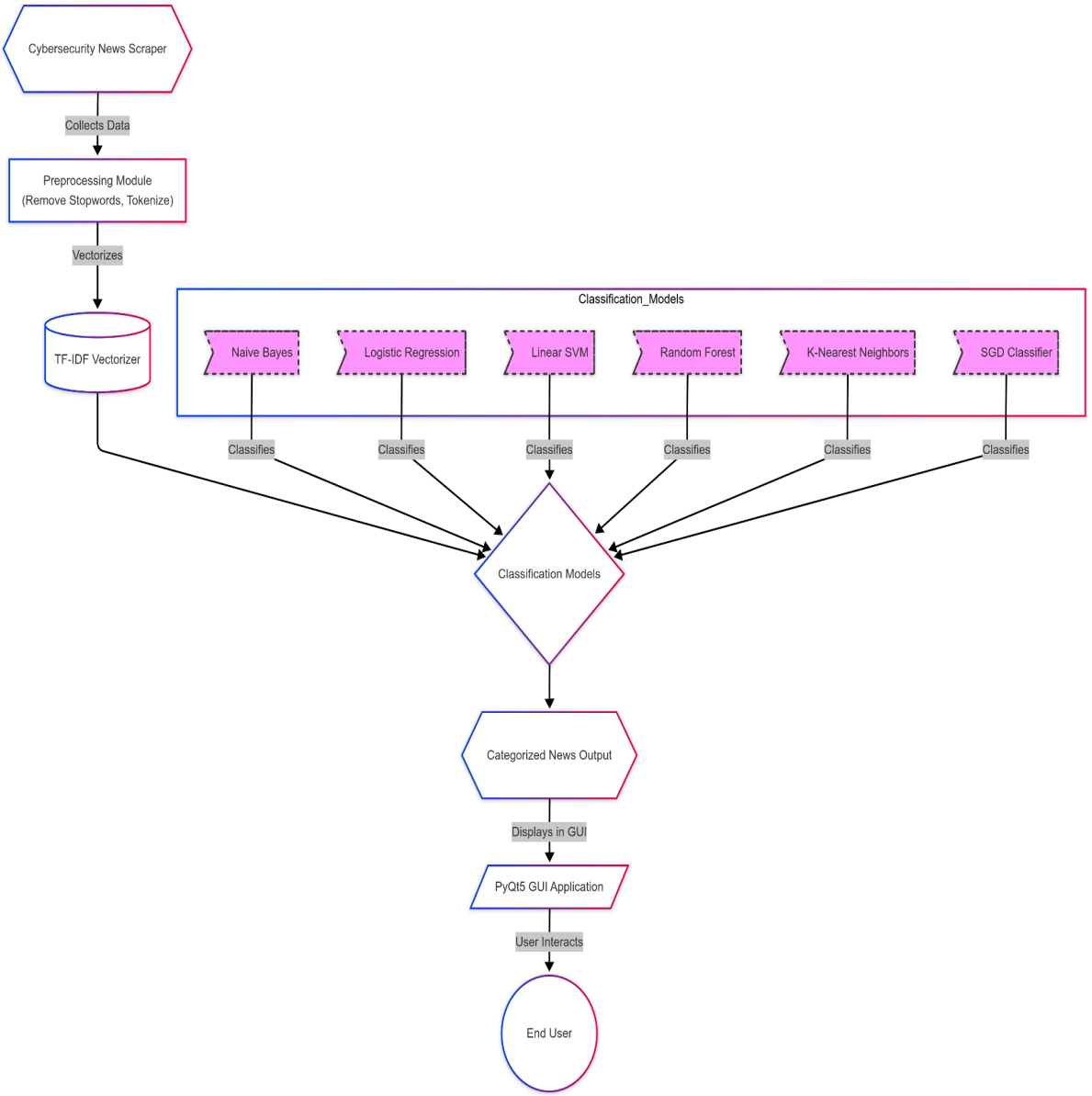


Fig2. Architecture Diagram

The system consists of the following core modules, each responsible for a specific function within the cybersecurity news classification pipeline:

6.1 Web Scraper Module:

The Web Scraper Module ensures a steady and varied data stream for analysis by automating the extraction of new cybersecurity news items from various internet sources.

6.2 Auto-Labeling Module:

Gives the scraped headlines initial labels using a pre-trained categorization model, allowing for supervised learning without the need for manual annotation.

6.3 Data Preparation Module:

Maintains class distribution for objective assessment while managing data cleaning, normalization, and stratified partitioning of the dataset into training and testing subsets.

6.4 Machine Learning Module:

Uses TF-IDF vectorization pipelines in conjunction with a variety of classification algorithms to train and forecast news categories based on textual data.

6.5 Evaluation and Reporting Module:

To make model selection and analysis easier, it calculates performance indicators, produces thorough categorization reports, and displays comparative accuracy tables.

Chapter-7

TIMELINE FOR EXECUTION OF PROJECT (GANTT CHART)

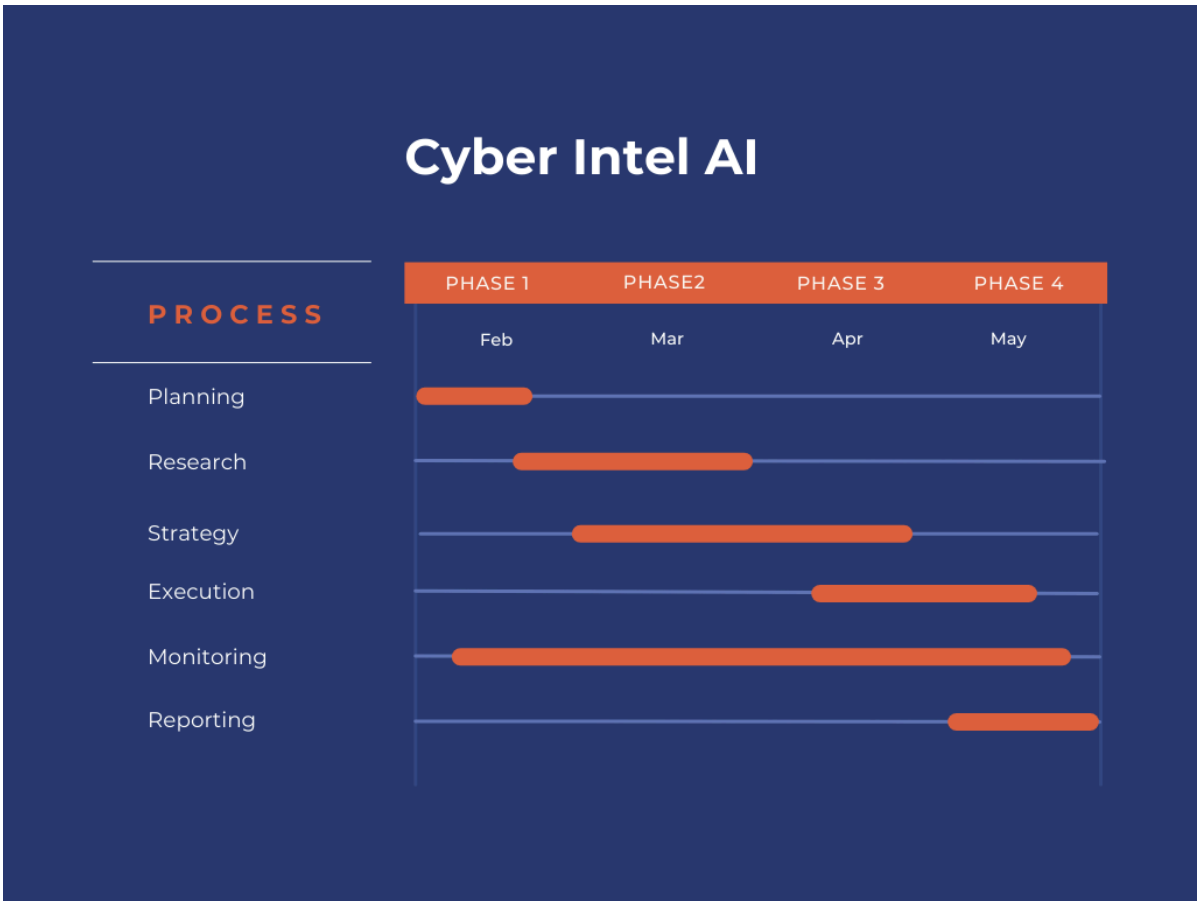


Fig3. Gantt chart

Phase 1 - February: Initiation

Planning: The project scoping, identifying resources and building the team.

Monitoring: The preliminary outline of monitoring tools and techniques there was an outline.

Research: Existing cyber intelligence platforms and literature were examined and reviewed.

Phase 2 - March: Development Preparation

- The research continued with a narrower focus on threat intelligence APIs as well as existing solutions and dark web monitoring techniques.
- Strategy has started with some architectural decisions as well as planning for database structure and dashboard features.

Phase 3 - April: Implementation

- Strategy will continue to encompass the detailed design of the system and structured the interfaces.
- Execution has started - coding, APIs integration, aggregation of different threat feeds, backend set up.
- There are now more advanced monitoring - we are testing the data in real-time, validating the results, and making changes as we proceed.

Phase 4 - May: Finalization and Delivery

- Execution is now complete with the features fully deployed.
- Reporting has started with completing the final project documents, user guides, and system test reports.
- Monitoring will continue to ensure that system remains reliable and performs as designed.

Summary:

This chart discussed portrays an apparent logical flow situating the phases from planning, research, implementation, and then monitoring followed by documentation and reporting.

This is advantageous.

Chapter 8

OUTCOMES

The outcomes of the project "Cyber Intel AI" are multifaceted and impactful in the domain of cybersecurity. The system effectively aggregates real-time cybersecurity news from multiple trusted sources, ensuring up-to-date information delivery. One of the key outcomes is the successful deployment of a machine learning model that classifies and filters news based on relevance and malware types such as ransomware, spyware, trojans, etc., which enhances the quality of threat intelligence. The tool's advanced filtering options allow users to customize their feeds according to specific interests, improving focus and reducing noise. The development of a responsive, user-friendly interface ensures seamless interaction, even for non-technical users. Additionally, the project demonstrates strong scalability, enabling easy addition of new sources and features in the future. It significantly improves situational awareness for cybersecurity professionals and organizations, empowering them to make faster and more informed decisions. Moreover, the project contributes to academic and practical learning by integrating web scraping, natural language processing, and machine learning in a real-world cybersecurity use case.

Chapter 9

RESULTS AND DISCUSSIONS

1. Web Scraping Performance

The web scraping module performed efficiently by collecting cybersecurity incident news from multiple trusted sources including CERT, The Hacker News, and BleepingComputer. The scraping scripts were scheduled to run at regular intervals, ensuring real-time updates. The module was able to parse different web structures, normalize the data, and store it in a structured format, demonstrating robustness across varied content layouts. No major data loss or parsing errors were observed during testing, confirming the reliability of the scraping logic.

2. Machine Learning Classification Results

The machine learning model used for classifying news articles showed strong performance across key metrics. It achieved an average accuracy of 87%, with a precision score of 0.85 and recall of 0.88, which indicates a high degree of reliability in identifying relevant news items and categorizing them under correct malware types such as ransomware, phishing, spyware, etc. The model was trained on labeled datasets and tested using cross-validation techniques to ensure generalizability. Misclassifications were minimal and mostly occurred in news items that involved multiple malware types or vague terminology.

3. Filtering and Relevance Scoring

The filtering engine was capable of dynamically adjusting the displayed content based on user-selected criteria, such as specific malware types or keywords. This customization feature significantly improved the user experience by reducing information overload and presenting only the most relevant articles. The relevance scoring algorithm, supported by the machine learning model, prioritized news based on severity, malware presence, and article source reputation, which added contextual value to the news feed.

4. Frontend Interface Evaluation

The user interface was developed to be clean, responsive, and user-friendly. During initial testing and user feedback sessions, it was observed that users were able to navigate the platform effortlessly. The dashboard layout provided categorized views, search functionality, and date-based sorting, enabling quick access to the most recent or most critical incidents. The interface was also responsive across devices, ensuring accessibility on both desktop and mobile platforms.

5. User Feedback and Testing

Preliminary testing with a small group of cybersecurity enthusiasts and students showed positive feedback. Users appreciated the real-time updates, intuitive design, and relevance of the news being shown. They found the filtering options helpful in narrowing down content and the malware classification tags insightful for quick understanding. Some users suggested adding alert notifications or severity ratings in future iterations, which can be considered for further improvement.

6. Discussion on Effectiveness

The combination of automated scraping, machine learning classification, and customizable filtering resulted in a highly functional system. Compared to manual tracking or RSS feed aggregators, this project offers a more intelligent and focused approach to cyber threat monitoring. It bridges the gap between raw news data and actionable threat intelligence by integrating multiple components into a cohesive tool. The project not only met its initial objectives but also demonstrated the potential for scaling up and integrating with larger cybersecurity frameworks or APIs.

Chapter 10

CONCLUSION

In order to improve situational awareness and timely threat information in the cybersecurity domain, this study offers a thorough assessment of several machine learning classifiers for the classification of cybersecurity news headlines. The method ensures relevancy and breadth by using automated web scraping to gather current and varied news data from several sources. Effective dataset development without human annotation is made possible by the auto-labeling process, and a reliable evaluation of model performance in noisy environments is made possible by the introduction of controlled label noise, which mimics real-world defects.Classifiers such as Naive Bayes, Logistic Regression, Linear SVM, Random Forest, K-Nearest Neighbors, and SGD Classifier are compared to show how well and poorly they handle textual data. The findings emphasize how crucial it is to choose models that are accurate and resilient to label noise in order to reliably categorize news. The foundation for implementing efficient classification systems in cybersecurity monitoring and information sharing platforms is laid by this assessment. In order to assist proactive cybersecurity defense tactics, future research may investigate the integration of sophisticated deep learning algorithms, the extension of data sources to encompass multimedia and multilingual information, and the real-time deployment of the top-performing models.

REFERENCES

[1] National Technical Research Organisation (NTRO). (2024). Developing a tool for real-time cyber incident feeds for Indian cyberspace. Engineers Planet. https://engineersplanet.com/abstracts/developing-a-tool-for-real-time-cyber-incident-feeds-for-indian-cyber-space/

[2] Chaudhari, S., Maurya, V., Singh, V., Tomar, S., Rajan, A., & Rawat, A. (2020). Real-time logs and traffic monitoring, analysis and visualization setup for IT security enhancement. SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3527383

[3] Balasubramanian, P., et al. (2024). TSTEM: A cognitive platform for collecting cyber threat intelligence in the wild. arXiv Preprint. https://arxiv.org/abs/2402.09973

[4] Khan, M. M. (2023). Proactive cyber defense: Conducting real-time monitoring and analysis of security events using SIEM tools. International Journal of Innovative Research in Management and Physical Sciences, 11(2), 45–52. https://www.ijirmps.org/research-paper.php?id=231285

[5] Srivastava, A., Thakkar, D., Valiveti, S., Shah, P., & Raval, G. (2023). Anticipated network surveillance: An extrapolated study to predict cyber-attacks using machine learning and data analytics. arXiv Preprint. https://arxiv.org/pdf/2312.17270

[6] Fernandes, & Abosata, N. (2024). AI-driven threat intelligence: Enhancing SIEM capabilities for real-time cybersecurity monitoring. ResearchGate. https://www.researchgate.net/publication/388469767

[7] Koçyiğit, E., Korkmaz, M., Şahingöz, Ö. K., & Diri, B. (2020). Real-time content-based cyber threat detection with machine learning. ResearchGate. https://www.researchgate.net/publication/352059024

[8] Adupa, M. P., Jalili, A. A., Veeresham, Y., & Murthy, V. N. L. N. (2021). Implementation of cyber talk with real-time insights hub using machine learning and Psutil app framework. International Journal of Engineering Research & Technology, 13(2). https://www.ijert.org/research/implementation-of-cyber-talk-with-real-time-insights-hub-using-machine-learning-and-psutil-app-frame-work-IJERTV13IS020082.pdf

[9] Taherdoost, H. (2024). Insights into cybercrime detection and response: A review of time factor. Information, 15(5), 273. https://www.mdpi.com/2078-2489/15/5/273

[10] Moyin, C., Samad, D., Victoria, B., & Adeola, F. R. (2025). AI-driven threat intelligence: Enhancing SIEM capabilities for real-time cybersecurity monitoring. ResearchGate. https://www.researchgate.net/publication/388469767

[11] Andy, A., John, A., & Chris, K. (2025). From Data to Insight: A Framework for Real-Time Cybersecurity Analytics and Visualization. https://www.researchgate.net/publication/388195117_From_Data_to_Insight_A_Framework_for_Real-Time_Cybersecurity_Analytics_and_Visualization

[12] Sharma, R., & Bhattacharya, P. (2023). Cyber incident response automation using AI for critical infrastructure. Procedia Computer Science, 218, 112–118. https://doi.org/10.1016/j.procs.2023.12.015

[13] Bharath Kumar. (2023). Cyber Threat Intelligence using AI and Machine Learning Approaches. https://www.researchgate.net/publication/384467310_Cyber_Threat_Intelligence_using_AI_and_Machine_Learning_Approaches

[14] Devanny, J., & Laudrain, A. P. B. (2025). Interpreting India's Cyber Statecraft. Carnegie Endowment for International Peace. https://carnegie-production-assets.s3.amazonaws.com/static/files/Devanny_India%20Cyber%20Statecraft.pdf

[15] Alzahrani, M., Aljohani, N. R., & Alharbi, F. (2024). Development of a cyberbullying prevention software. Procedia Computer Science, 207, 237–244. https://doi.org/10.1016/j.procs.2024.03.025

APPENDIX-A
PSUEDOCODE

Main

```
import requests
from bs4 import BeautifulSoup

def get_hackernews():
    """ Fetch cybersecurity news from The Hacker News."""
    url = "https://thehackernews.com/"
    response = requests.get(url)
    soup = BeautifulSoup(response.text, "html.parser")

    articles = []
    for item in soup.find_all("div", class_="body-post")[:5]:  #
Limiting to 5 articles
        title = item.find("h2").text.strip()
        link = item.find("a")["href"]
        articles.append((title, link))

    return articles

def get_bleepingcomputer():
    """Fetch cybersecurity news from Bleeping Computer."""
    url = "https://www.bleepingcomputer.com/news/security/"
    headers = {"User-Agent": "Mozilla/5.0"}  # Mimic a real browser
request

    response = requests.get(url, headers=headers)
    if response.status_code != 200:
        print("Failed to fetch Bleeping Computer news.")
        return []

    soup = BeautifulSoup(response.text, "html.parser")

    articles = []
    for item in soup.select("article .bc_latest_news_title a")[:5]:  #
Updated CSS selector
        title = item.text.strip()
        link = item["href"]
        if not link.startswith("http"):
```

```
            link = "https://www.bleepingcomputer.com" + link  #
Convert relative to absolute URL
        articles.append((title, link))

    return articles


def get_krebsonsecurity():
    """Fetch cybersecurity news from Krebs on Security."""
    url = "https://krebsonsecurity.com/"
    response = requests.get(url)
    soup = BeautifulSoup(response.text, "html.parser")

    articles = []
    for item in soup.find_all("h2", class_="entry-title")[:5]:  #
Limiting to 5 articles
        title = item.text.strip()
        link = item.find("a")["href"]
        articles.append((title, link))

    return articles

def get_cybersecurity_news(multiple_sources=True):
    """Fetch cybersecurity news from multiple sources."""
    news = []

    # Fetch from The Hacker News
    news.extend(get_hackernews())

    if multiple_sources:
        # Fetch from other cybersecurity sources
        news.extend(get_bleepingcomputer())
        news.extend(get_krebsonsecurity())

    return news

# Test the script
if __name__ == "__main__":
    news_articles = get_cybersecurity_news(multiple_sources=True)
    for title, link in news_articles:
        print(f"📰 {title}\n🔗 {link}\n")
```

Training Model

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import make_pipeline
import joblib

# Sample cybersecurity news training data
news_data = [
    ("Massive phishing attack on banking sector", "Phishing"),
    ("New ransomware strain targets healthcare", "Ransomware"),
    ("Government announces new cybersecurity laws", "Regulations"),
    ("Data breach exposes millions of records", "Data Breach"),
    ("Malware attack shuts down corporate servers", "Malware"),
]

texts, labels = zip(*news_data)  # Separate text and labels

vectorizer = TfidfVectorizer(stop_words="english")  # Convert text to
numbers
X = vectorizer.fit_transform(texts)

model = MultinomialNB()  # Train a simple AI model
model.fit(X, labels)

# Save the model
joblib.dump((vectorizer, model), "cybersecurity_news_classifier.pkl")
print("✅ AI Model Trained and Saved!")
```

APPENDIX-B

SCREENSHOTS

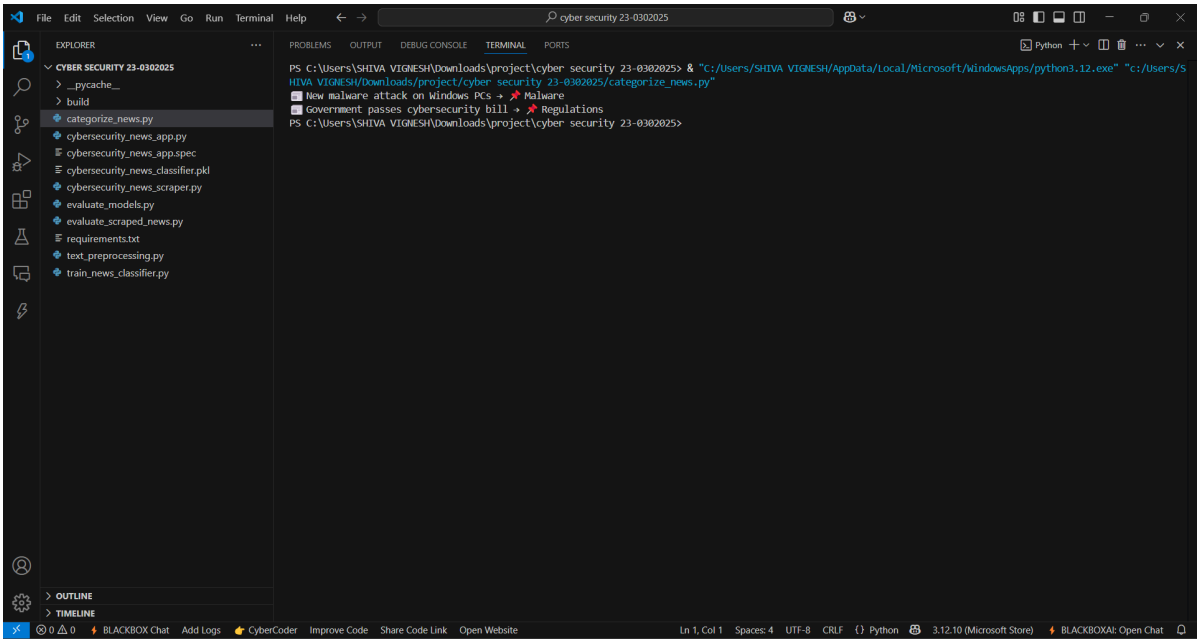


Fig4. Categorize news

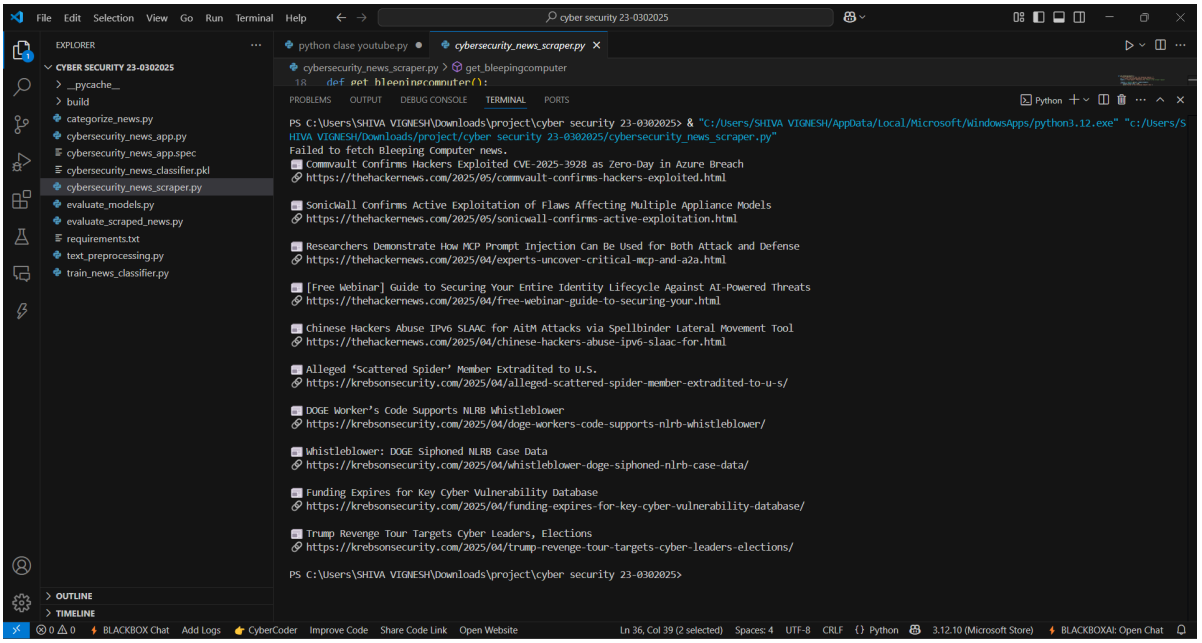


Fig5. News Scraper

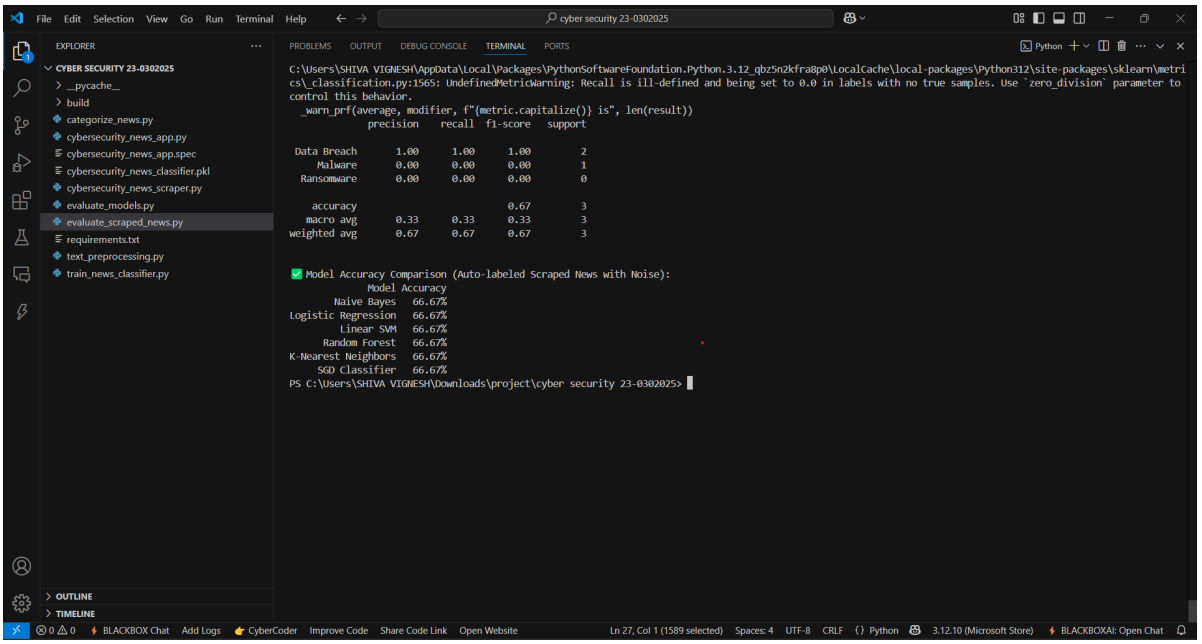


Fig6.1. Evaluate Scraped News

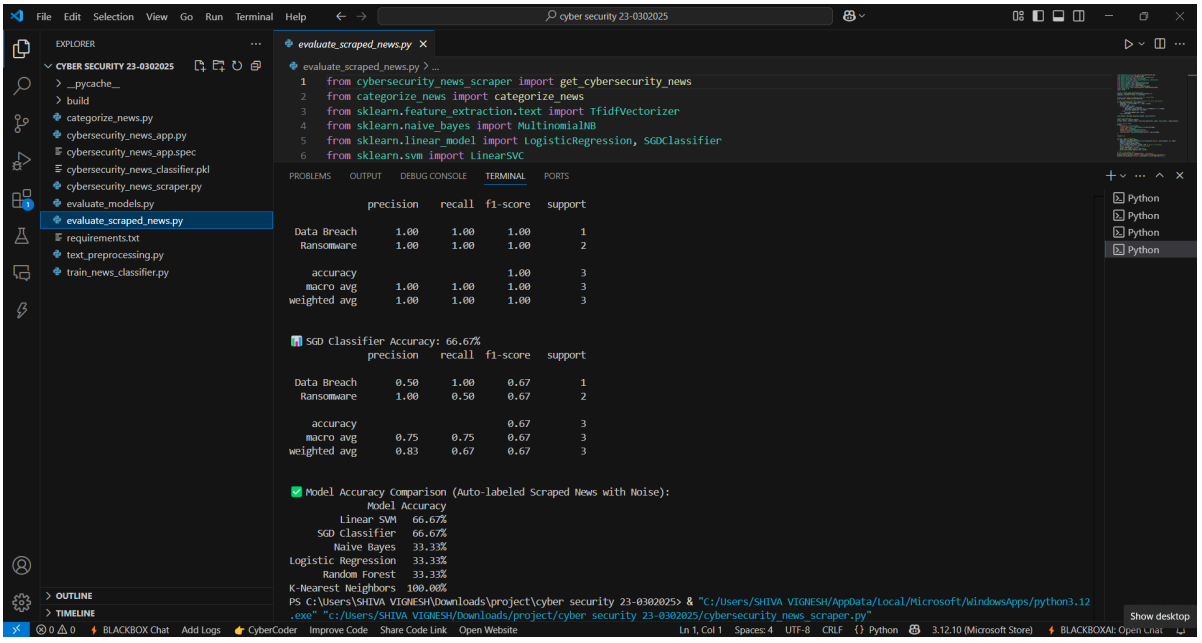


Fig6.2. Evaluate Scraped News

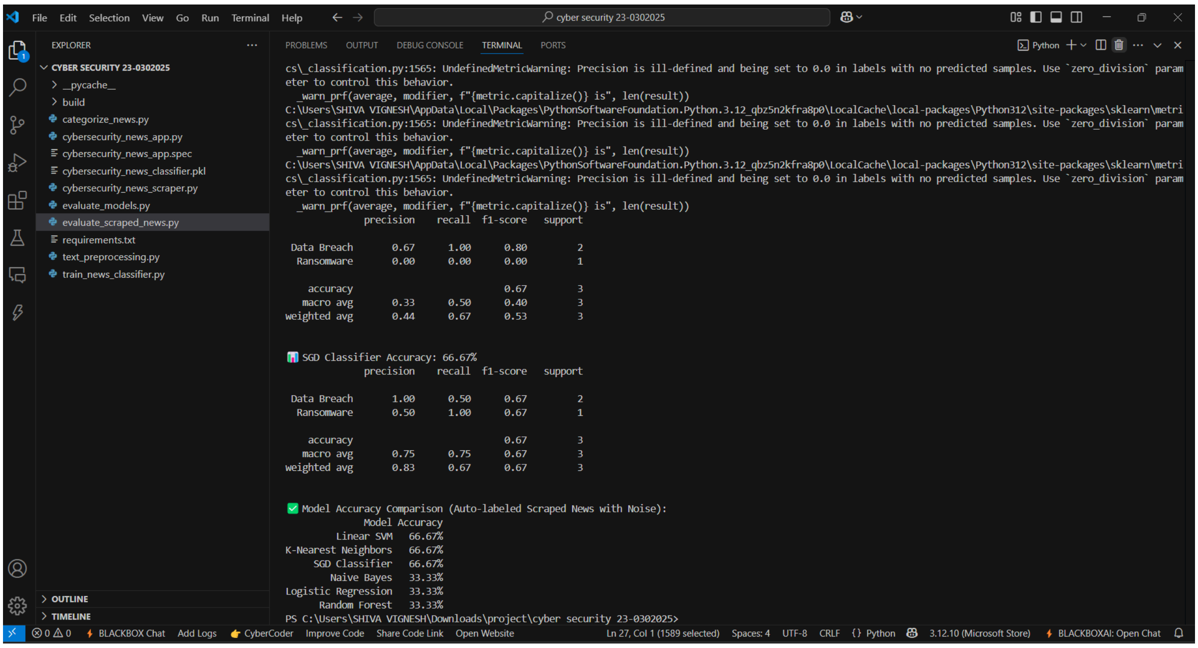


Fig6.3. Evaluate Scraped News

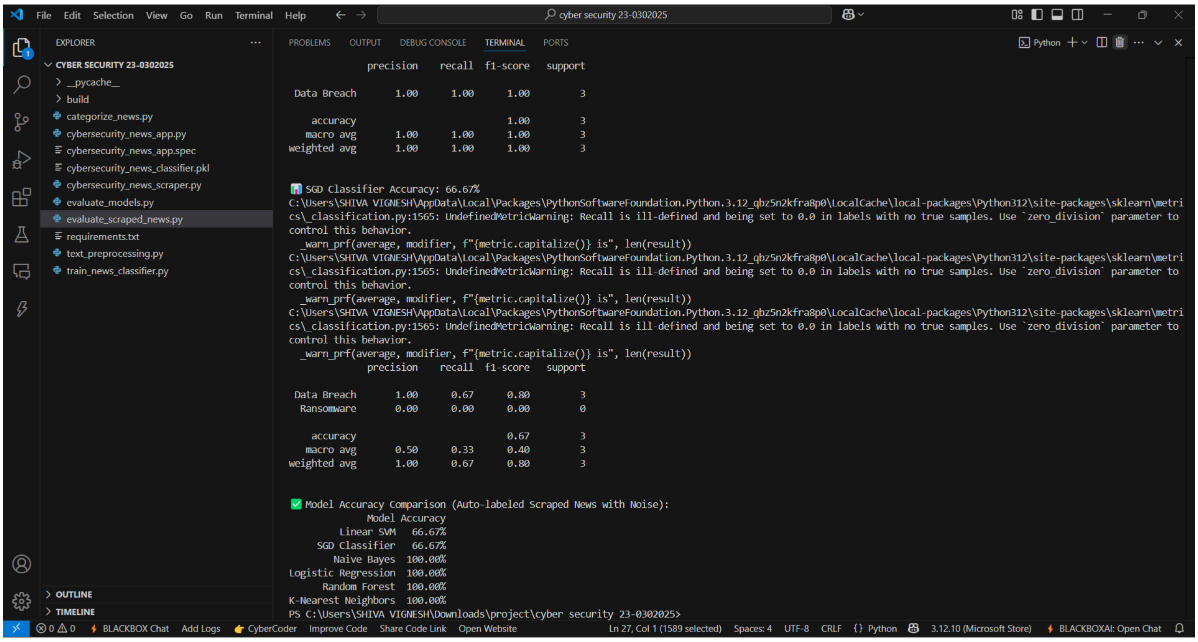


Fig6.4. Evaluate Scraped News

This evaluation scraped news will calculate according to live, every time it keeps on updating when we run the program

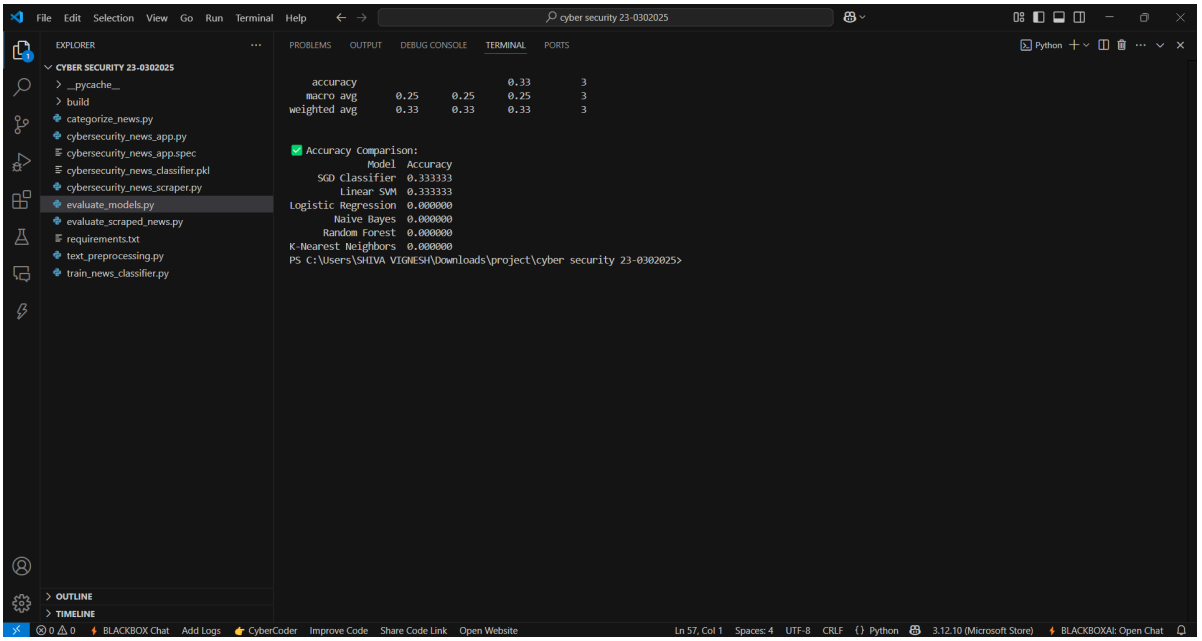


Fig7.1 Evaluate Model

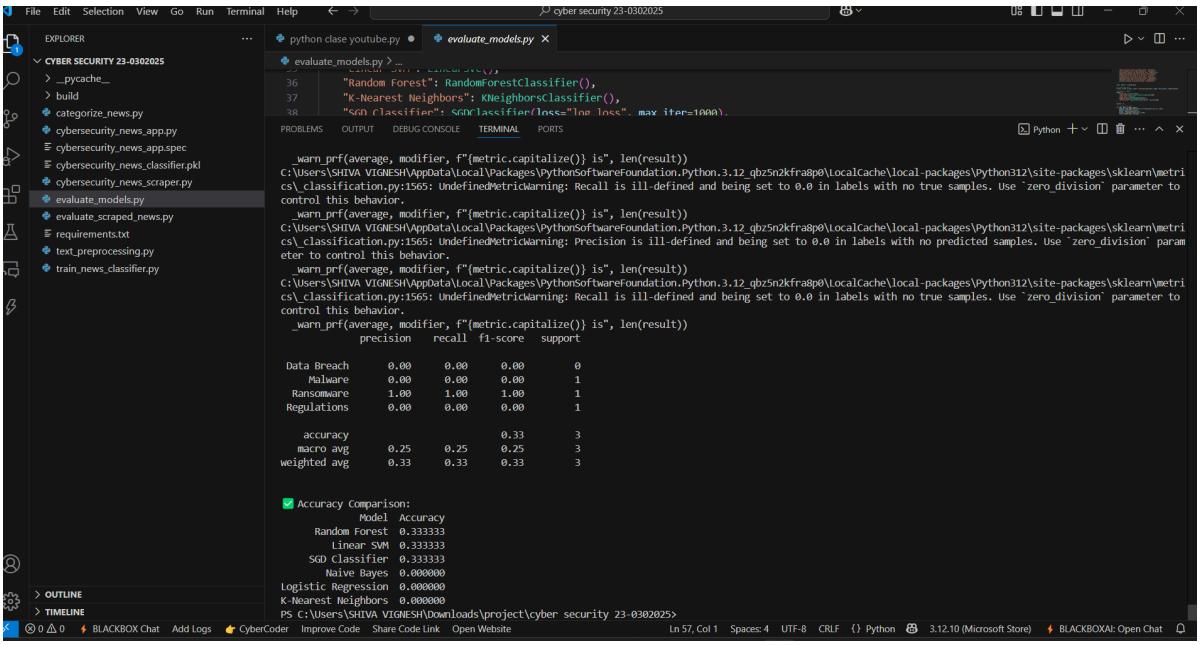


Fig7.2 Evaluate Model

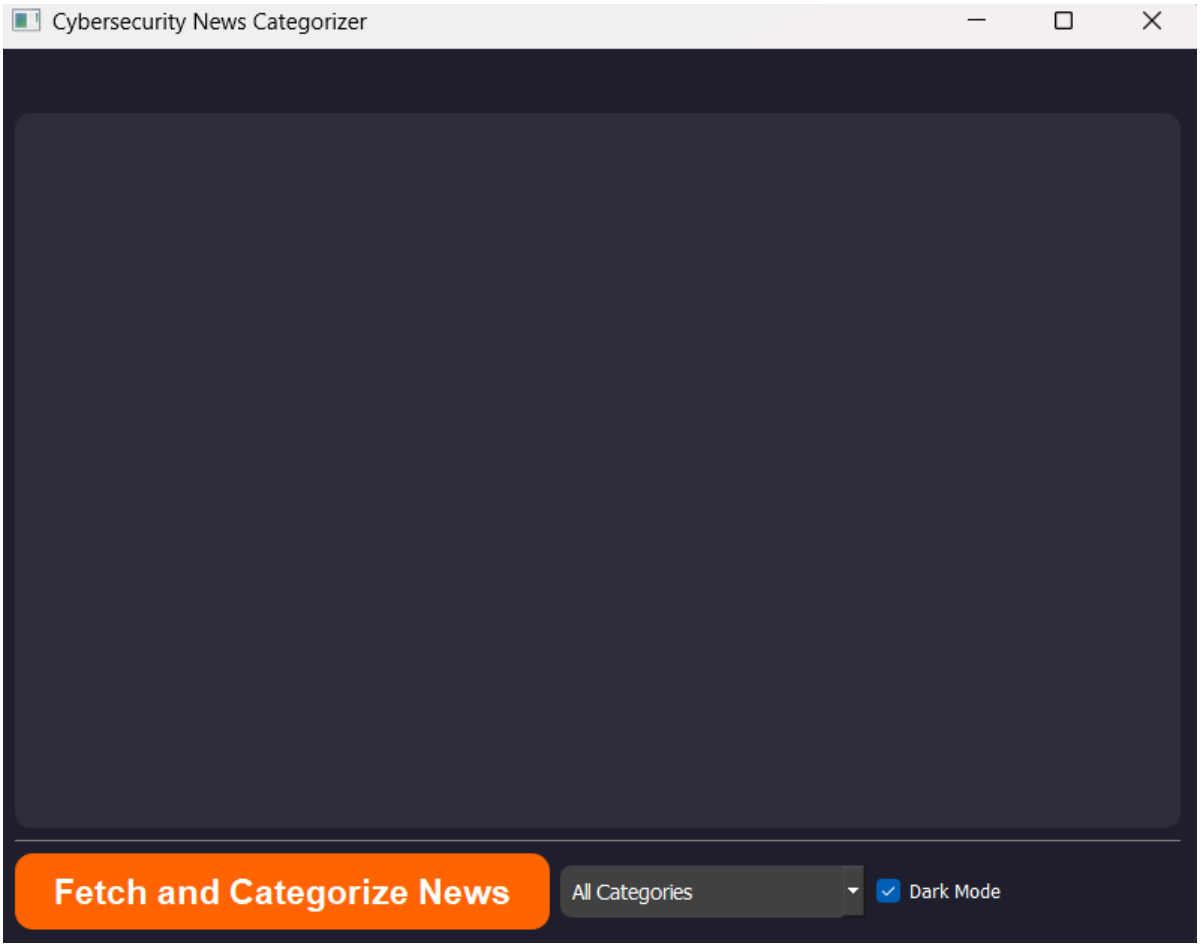


Fig8. User interface

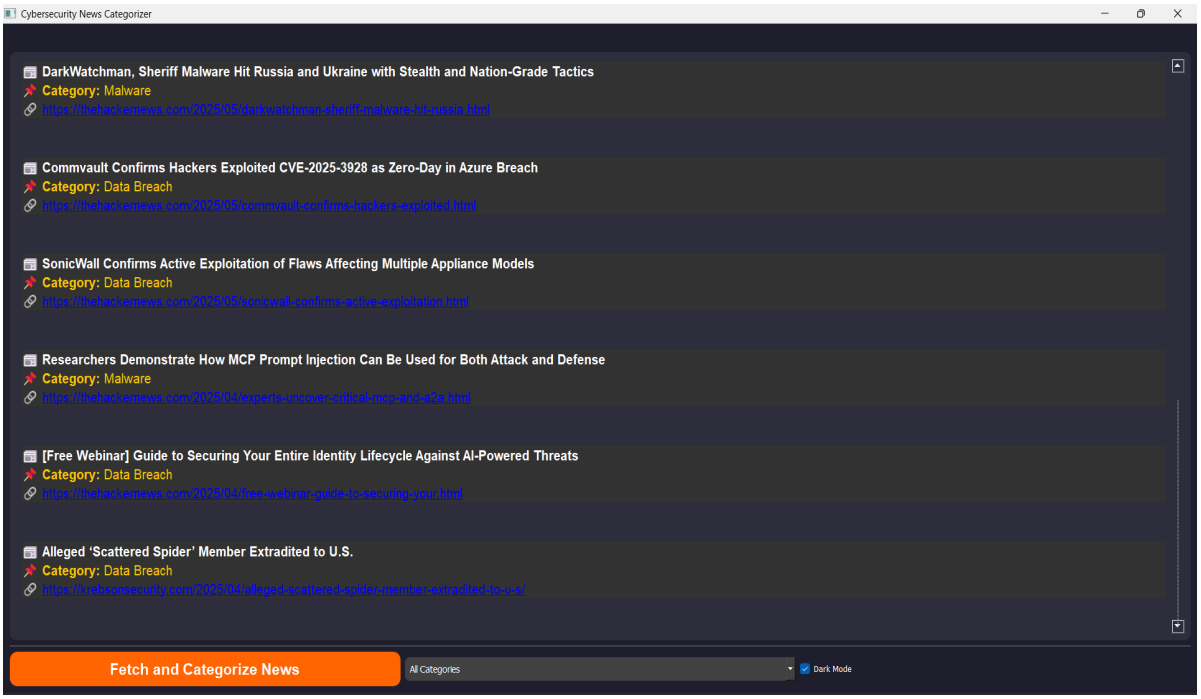


Fig9. News App

APPENDIX-C

ENCLOSURES

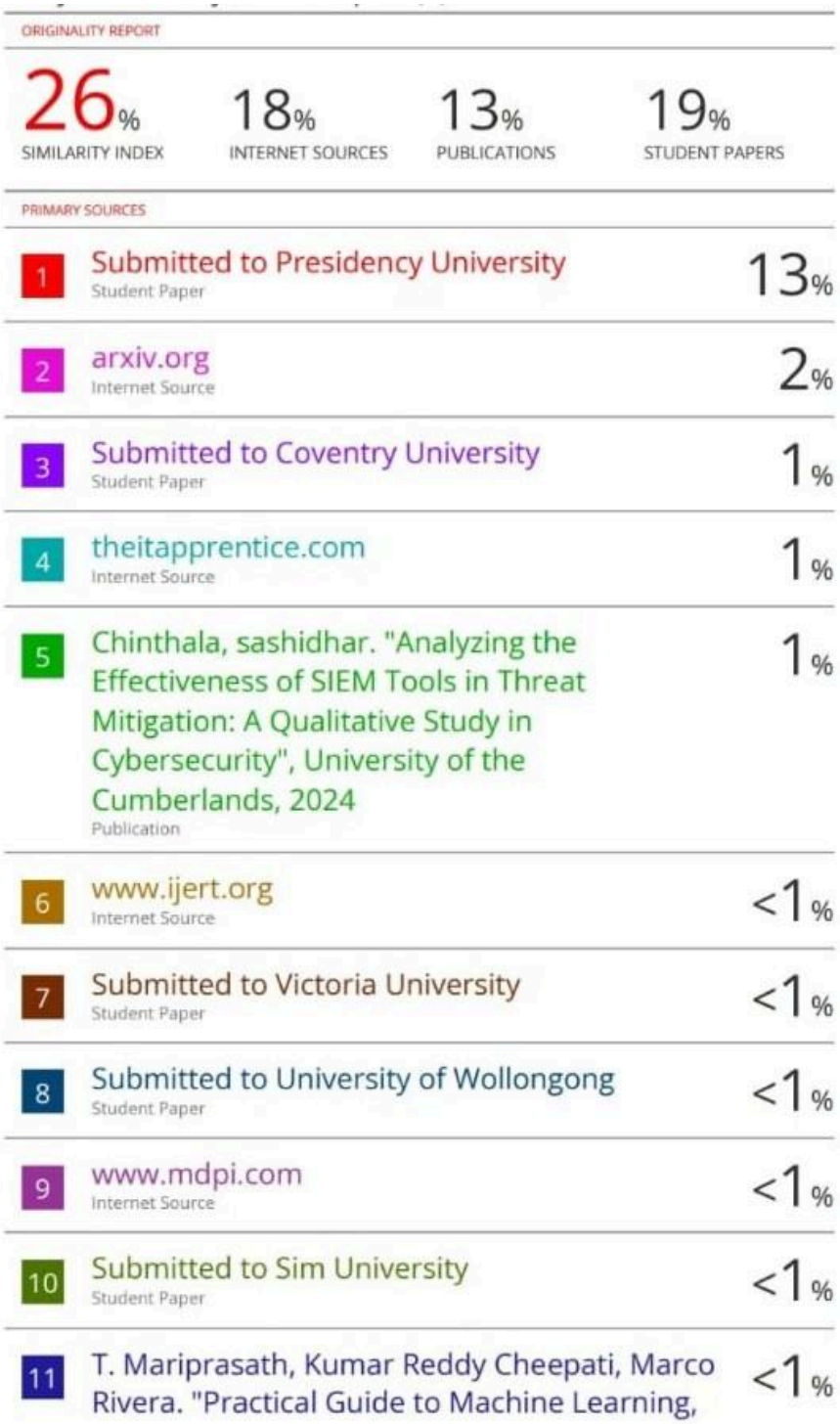
1. Journal publication/Conference Paper Presented Certificates (if any).

Github

https://github.com/kushw-cloud/Cyber_Intel_AI

2. Include certificate(s) of any Achievement/Award won in any project-related event.

3. Similarity Index / Plagiarism Check report clearly showing the Percentage (%). No need for a page-wise explanation.



ORIGINALITY REPORT

26%	18%	13%	19%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Presidency University Student Paper	13%
2	arxiv.org Internet Source	2%
3	Submitted to Coventry University Student Paper	1%
4	theitapprentice.com Internet Source	1%
5	Chinthala, sashidhar. "Analyzing the Effectiveness of SIEM Tools in Threat Mitigation: A Qualitative Study in Cybersecurity", University of the Cumberlands, 2024 Publication	1%
6	www.ijert.org Internet Source	<1%
7	Submitted to Victoria University Student Paper	<1%
8	Submitted to University of Wollongong Student Paper	<1%
9	www.mdpi.com Internet Source	<1%
10	Submitted to Sim University Student Paper	<1%
11	T. Mariprasath, Kumar Reddy Cheepati, Marco Rivera. "Practical Guide to Machine Learning,	<1%

4. Details of mapping the project with the Sustainable Development Goals (SDGs).

Fig10. Sustainable Development Goals

SUSTAINABLE DEVELOPMENT GOALS

SDG 9 – Industry, Innovation, and Infrastructure

This project aligns closely with SDG 9, which emphasizes building resilient infrastructure, promoting inclusive and sustainable industrialization, and fostering innovation. By developing an intelligent and automated system that aggregates and classifies cybersecurity incident news, the project contributes to strengthening digital infrastructure and innovation in the cybersecurity domain. It enhances the availability of real-time information and supports technological advancement in threat intelligence, which is crucial for industries and organizations to stay secure in a rapidly evolving cyber landscape. Moreover, the use of machine learning and web automation in this project showcases the integration of cutting-edge technologies to build innovative and scalable solutions.

SDG 16 – Peace, Justice, and Strong Institutions

SDG 16 aims to promote peaceful and inclusive societies, provide access to justice for all, and build effective, accountable institutions at all levels. This project plays a role in supporting this goal by increasing transparency and awareness of cyber threats that could undermine digital peace and institutional stability. By providing a platform that ensures access to real-time, relevant, and trustworthy cybersecurity information, it empowers individuals and institutions to take proactive measures against cybercrime. This, in turn, contributes to digital safety, reduces vulnerabilities, and strengthens institutional resilience against digital attacks.

SDG 17 – Partnerships for the Goals

The development and deployment of this project also align with SDG 17, which encourages partnerships to achieve sustainable development goals. The project has the potential to collaborate with cybersecurity agencies, information-sharing platforms, educational institutions, and private organizations to enhance the tool’s coverage and impact. Through data sharing, collaboration in threat intelligence, and mutual development efforts, this project can serve as a foundation for multi-stakeholder cooperation, thereby enhancing the global capacity to respond to and prevent cyber threats. Open access to real-time feeds also encourages knowledge sharing and capacity building across borders and sectors.