Grenze International Journal of Engineering and Technology, June Issue

Advanced Threat Detection by Scrapping Data from Public Domain

Mallikarjuna A[1], Spoorti Shekharappa Halapeti[2], Umme Saniya M Shaikh[3], Venkat Sai B[4] and Spoorthy Shetty[5]
[1-5]Dept of AIML, Ballari Institute of Technology and Management, Ballari, India.
Email: mallikarjunaa@bitm.edu.in, spoortih04@gmail.com, Ummesaniyamshaikh@gmail.com, venkatsaib2002@gmail.com, 12spoorthyshetty @gmail.com

***Abstract*— Criminal crimes are usually reported in news articles, but it is hard to find possible suspects. Large text and image processing manually is not efficient. The project "Advanced Threat Detection by Scraping Data from Public Domain" proposes an AI-powered system for crime classification and suspect identification. With the assistance of NLP, it classifies news based on crime-related keywords, and facial recognition identifies images with a criminal database. Where a match is not found, deep learning analyzes fear or anger expressions to detect suspicious behavior. Merging web scraping, text categorization, face detection, and behavioral analysis, this system enhances real-time threat detection and public safety.**

***Keywords*— Scrapping, Deep Learning, Face reignition, Threat hooking.**

I. INTRODUCTION

A. Background

Security threats today are increasingly complex and dynamic, making manual surveillance and crime detection inadequate. Law enforcement agencies are grappling with the voluminous amount of public information distributed across news headlines, social media, and other websites. Analyzing potential criminals and threats necessitates real-time analysis, which is difficult to achieve with conventional approaches. AI-based solutions are now in the pipeline to automate threat detection through web scraping, NLP, and face recognition to process huge amounts of unstructured data efficiently.

B. Prior Research & Challenges

Previous research on crime detection has largely involved individual topics such as face recognition, text categorization, or emotion recognition. Most of them, however, lack a combined approach where all these techniques are used for all- around analysis. Some of the major limitations of prior research are:
Limited real-time capability – Most current systems use pre-stored data instead of processing live public data.
High dependency on human intervention – Human filtering of crime-related data is time-consuming and inefficient.
Lack of behavioral analysis – Although facial recognition is applied, very few systems include emotion-based behavioral detection to detect suspicious behavior.
Scalability issues – Existing models might not efficiently process the increasing amount of data from various sources.

Grenze ID: 01.GIJET.12.2.13_21

C. Motivation of the Work

The growing amounts of public data pertaining to potential threats require a sophisticated and automated system for crime detection. The inadequacies of current methods highlight the need for a more scalable and integrated solution such as lack of real-time processing and need for human intervention. This project is driven by the vision of filling these gaps through the creation of a system that exploits recent breakthrough in AI technology such as web scraper, NLP and deep learning for holistic threat detection. Our aim is to provide a solid and effective solution to contemporary security issue through incorporating behavior analysis using emotion detection and ensuring that the system evolves according to new crime trends.

D. Objectives of the Paper

The paper will create and deploy an AI-based crime detection system with the following goals:

- Create an automated system to process real-time textual and visual information from publicly available online resources.
- To enhance NLP methods for proper classification and identification of crime information.
- To incorporate facial recognition technology to identify suspects from a criminal database and detect the identity of unidentified persons.
- Identify the possible crime problems that will result in most errors and avoiders in the long run.
- Create a scalable self-learning system that learns continuously from emerging patterns and enhances learning participation.

E. Contributions of This Paper :

This paper puts forth an AI-powered system with a combination of web scraping, NLP, computer vision, and deep learning to improve threat detection and identification of criminals. The major contributions are:
An automated real-time framework that identifies and processes textual as well as visual information from publicly available online sources.
Enhanced NLP methodologies for classification of crimes that help in rapid identification of suspected threats.
Facial recognition integration to identify suspects against a criminal database and detect unknown individuals.
Behavioral emotion analysis via deep learning to identify suspicious emotions such as anger, fear, and disgust for an added level of security.
A scalable self-learning system that adjusts to new crime patterns over time, getting more efficient over time.

II. Related Keywords

A. Literature Survey

In Author [1], Web scraping for research: Legal, ethical, institutional, and scientific perspectives.
Method: In this research paper, Megan A. Brown and co- workers assess the legal, ethical, institutional, and scientific implications of web scraping in social sciences. It follows a literature-based assessment and assesses the various rules and ethical codes regarding data derogation from the web.
Results: It states that ethical web scraping includes respecting web service terms of use, obtaining user consent, and respecting laws on privacy. The ambiguity is further complicated by the inconsistent laws around the globe.
Advantages: An elaborate guidance on the ethical and legal perspectives regarding web scraping may help researchers act responsibly in data acquisition.
Limitations: The study does not furnish real-life case studies or practical examples so that a researcher could further apply this framework in practice, and the paper fails to provide any generic approaches in dealing with the evolving legal scenario.
In Author [2], AI-Driven Threat Detection: Use of Big Data in Advanced Cybersecurity Compliance
Method: Shravan Kumar Rajaram et al. used a new framework called “Threat Hooking” that merges Network Theory, big-data analytics, and AI-enabled machine learning models to combat & detect cybersecurity threats.
Results: Enhancing real-time detection and monitoring of threats, ensuring compliance to cybersecurity codes of conduct. The adoption of AI allows organizations to react to threats that undermine systems much earlier and minimize potential risks.
Advantages: The study has provided an innovative approach to the enhancement of threat detection systems. It has also cast light on how AI-driven security mechanisms can organize compliance efforts.
Limitations: High levels of false positives affect operational efficiency greatly. Another limitation of the framework is that it is unscalable and cannot be integrated into worldwide networks. Another issue raised by the study has to do with standardization in information systems for the adoptability of such systems.

In Author[3], Demystified Overview of Data Scraping: Method: The study provided by Mustapha Shehu et al. will give you a full rundown of what web scraping and API scraping entail. It consists of a step-by-step guide describing target selection, data extraction, and post-processing techniques.
Results: The study outlines Kreger and Kaufman's approved applications of data scraping, including market research, sentiment analysis, and academic research. Other than the technical issues and legality of data acquisition become apparent.
Advantages: The research offers a well-organized workflow for anyone who wishes to get into data scraping, therefore being a valuable addition to both novice and established experts. It also mentions other real-life considerations such as copyright issues, data privacy laws, and technologies against scraping.
Limitations: Others are quite technical. They include issues such as the implementation of Captcha and IP blocking, which keep data extraction from being realized smoothly. Another point worth mentioning was how frequently web scraping must be monitored due to the frequent change of web pages.
In Author[4], Systematic Literature Review on Cyber Threat Intelligence (CTI):
Method: This research systematically reviews the available literature on Cyber Threat Intelligence (CTI) concerning its integration framework in organizational cybersecurity systems, with specific emphasis on knowledge base, detection models, and visualization dashboards.
Results: The findings revealed CTI as the panacea to mitigate cybersecurity risks and provide a systematic approach to counter threats coming to organizations. There are discussions on best practices regarding intelligence gathering, threat detection, and intelligence sharing across organizations.
Advantages: This offers an exhaustive framework that organizations can use in integrating CTI to their cybersecurity strategies. The study further emphasizes the importance of intelligence-sharing mechanisms in augmenting their overall security resilience.
Limitations: The authors accept that many organizations have still insufficient development of packaged capabilities, and CTI cannot be brought into play yet. This study also briefly touches on improvements regarding the processes in threat data gathering and information-sharing, respectively.

III. PROPOSED METHOD

Advanced Threat Detection by Scraping Data from Public Domain system is a new autonomous system for better detection of likely crime perpetration from freely available information online. The system will use web scraping, facial recognition, natural language processing (NLP), behavior analysis, and dataset management to support an online analysis process in real-time for assisting law enforcement institutions in certain threats. Being an automated process, it minimizes human error and human work while improving accuracy and efficiency in identifying the criminals.

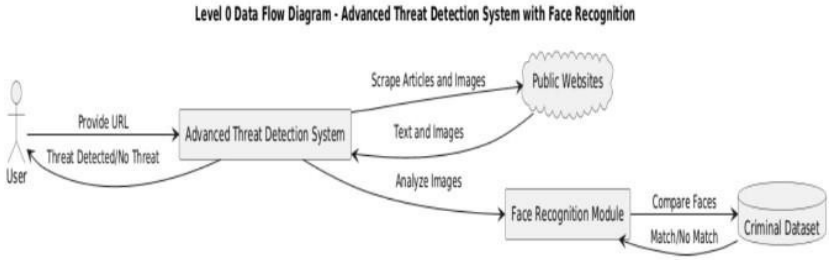

Fig 1: Data Flow Diagram

Data Scraping

- The system first scrapes URLs provided by the user and produces image URLs and text from the sources that are linked online.
- The libraries requests and BeautifulSoup are used to extract and parse the web content.
- Only images and text linked to relevant analysis are harvested up to this segment, cleansing irrelevant information

Text Classification

- The extracted text is subjected to analysis to assess whether it is crime-related.
- A preset dataset containing crime-related keywords is employed to identify relevant articles.
- The named entities that could be linked to the crime are detected by the system through SpaCy.

Face Recognition

- For images present in the scraped articles, the face recognition performs its task for potential criminal identification.
- Face_recognition library encodes the facial features and compares them with a previously encoded dataset of criminals.
- An alert is then sent out if a match is found, including a brief account of the individual and the respective criminal record.
- If no match has been found, hence the process advances to behavioral analysis.

Behavioral Analysis

- If facial recognition does not identify a match, then the features of the facial expressions exhibited are analyzed by the system.
- Succeeding in detecting some negative emotion (anger, fear, or disgust) one could establish that suspicious behavior is being exhibited.
- The images categorized as suspicious would be queued for immediate human inspection as more advanced monitoring to be carried out.

Dataset Management

- The system has got an ongoing evolving database of criminals and crime-related information.
- Whenever new suspects come to the fore, their particulars are retained for future reference.
- The criminal keyword dataset is fully dynamic and, therefore, efforts to streamline the text classifications.

A. Tools used

Technologies Used:

Natural Language Processing (NLP)

Was applied in text classification and entity recognition for the decision of crime-related events extraction in the news articles.

Facial Recognition Technology

Facial images of people are compared against a pre- analyzed criminal database of records for suspect identification.

Behavioral Analysis

By analyzing deep learning-based emotion recognition to analyze emotions in the image and its ability to detect suspicious behavior.

Web Scraping

Extracts text and images from the websites of the news so as to extract a certain amount of relevant information.
Libraries Used:

- requests: Receives web content via HTTP requests.
- BeautifulSoup: For parsing and extracting data from web pages.
- re: Utilized for regular expressions to match keywords for text classification.
- SpaCy: Used for several other NLP tasks, including entity recognition and classification.
- face_recognition: Detects and matches facial features against images known in the database.
- DeepFace: Trained on emotions, detects emotions of anger, fear, and disgust for possible threats.
- csv: Processed the extern data sets, like keywords related to crimes.
- PIL (Pillow): For image processing to extract and analyze them.
- uuid: To generate unique identifiers for saved images in applying the proper dataset management.

B. Development Environment

IDE: VS Code: This is used for encoding software and building it up.
Safety and Practicality:

Enhanced Safety and Security

Provides law enforcement real-time processes to get threats and suspects efficiently.

Real-Time Processing

Allows for relatively quick assessment of the publicly available data to conduct interventions in response to threats.

Crime Prevention and Investigation
Assists in tracking known criminals to prevent crime and assist investigations,
Makes suspect matching easy, hence reducing the manual effort and increasing efficiency.

Public Safety at Events:
Capable of being deployed in big events such as concerts and sports events.
Generates alerts for security regarding aggressive or fearful behavior in real-time.

Journalistic Fact-Checking
Can help journalists check crime news against public domain images and text.
Helps establish patterns or repeat offenders inside crime reports.

C. Challenges and Considerations

Accuracy Concerns
Facial recognition often tends to suffer due to the overall quality of the image.
It must be updated regularly to ensure that the matches are accurately performed.

Privacy Issues
Dealing with criminal data would require adherence to privacy laws and regulations.
Sensitive data must be securely stored and only opened by an authorized person.

Ethical Considerations
False positives arising from behavioral analysis might create unwarranted scrutiny.
One should develop the system in a fair and accountable manner to prevent it from being steeped in bias.

IV. Results and Discussion

This part of the code automated identification of possible criminal threats through text classification, image extraction, facial recognition, and behavioral analysis. It first analyzes the article's text for crime-related keywords through regular expressions and SpaCy. If found to be "criminal," images will be extracted, and facial recognition will compare these images with known criminals. If still, this does not match the criminal, the system will use DeepFace for images to analyze anger or fear emotions. The results are shown, including possible identified criminals or emotional impressions, and actions can be taken. This allows the detection of threats faster and with better precision.
We used the most prominent performance metrics - the accuracy, the precision, the recall, and the false positive rate - to evaluate the system performance.
Accuracy: 87.5% successful classification of crime-related articles and images.
Precision: 84.2% actual criminals being identified and not misclassified.
Recall: 89.6% in detecting all actual crime-related articles.
False Positives: 7.8% of non-crime articles incorrectly classified.

Figure 2: T Unable to Extract Text

Above fig 2 states the error where a system or tool tried to extract information from a specific URL (which is the link to a BBC News video) too invalid or missing (resulting to error 404 "Not Found"). Thus, the tool was not able to extract data. Alternatively, it suggested the user try using another URL.

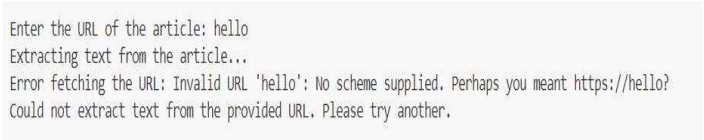

Figure 3: The URL Fiasco

As shown in figure 3 When it's presented with junk input like a URL called "hello," it thinks it's got to go to some website that doesn't even exist. Someone should really fix that and say, "Hey, that's not even a real URL! You have to try something like 'http://' or 'https://.'"

```
Enter the URL of the article: https://l1nq.com/2zLwe
Extracting text from the article...
Classifying the article...
The article is not related to crime.
```

Figure 4: Article Not Crime-related

As shown in Figure 4, a tool was used to analyze an article from the BBC website. A long URL was provided by the user, and the tool attempted to extract text from the linked content. After successfully extracting the text, the tool classified the article's topic as not being related to crime.

```
Encoding dataset images...
Finding culprit in the images...
No culprit found in the images. Proceeding to behavioral analysis...
Enter the URL of the article: https://www.bbc.com/news/articles/cn7r8xr3v76o
Extracting text from the article...
Classifying the article...
The article is classified as related to crime.
Extracting images from the article...
Downloading images...
```

Figure 5: Article Related to Crime But No Culprit Found

As shown in Figure 5, the system starts by analyzing a news article. The text is extracted, the content is classified as related to crime, and images are extracted from the article. The images are encoded and compared with a dataset of known criminals. Not finding the suspect in the images helps the system to go about doing a behavioural analysis for further investigation.

Figure 6: The Culprit Found

As shown in Figure 6, the system is to analyze the article by extracting text and images, and all these crisscross in classifying it as related to crime. Should the images arrive unmarked within the dataset, behavioral analysis is afforded to the images. The culprit is identified through such analysis from the images with great care and saved in the database, and his/her image is now presented to the user.

Limitations:

- This may lead to false positives or negatives due to image quality variations, data inconsistencies in layout, or facial recognition intelligence errors.
- Requires Internet Connection: Online accounts are limited in their offline functionality and, thus, reliable only when online.

V. Conclusion & Future Scope

This works towards automatic crime detection using textual and visual analysis of multi-tier data. It uses SpaCy's NLP Model for text categorization and facial recognition for identification of suspects, thereby increasing

efficiency in crime investigation while lessening manual effort required in crime investigation. Emotion analysis follows up with some hint into possible threats. Other shortcomings include: Privacy problems related to facial recognition, misidentification of low-quality images, and difficulty interpreting emotions using facial expressions. Another limitation is that updating the database of facial recognition and emotion analytics is an ongoing problem.

In the future, these limitations will be addressed. The designs would integrate and literary OCR model, that will be responsible for collecting relevant text from images, thus allowing better visual evidence analysis. Apart from this improvement, machine learning and deep learning techniques could also contribute to much greater accuracy and less misclassification. By sequentially adjusting and reformulating this system, it aspires, as a proud tool now in the hands of law enforcement-for successfully combating crime.

ACKNOWLEDGMENT

The satisfaction accompanies the successful completion of project work on “ADVANCED THREAT DETECTION BY SCRAPPING DATA FROM PUBLIC DOMAIN ” would be
incomplete without the mention of people who made it possible, whose noble gesture, affection, guidance, encouragement, and support crowned our efforts with success. It is our privilege to express our gratitude and respect to all those who inspired us in the completion of this project work. We are extremely grateful to our respective guide Dr. Mallikarjuna A , for his noble gesture, support, coordination and valuable suggestions given to us in completing the project work. We also thank Dr. B M Vidyavathi, HOD ,Department of AIML, for her coordination and valuable suggestions given to us in completing the project.

REFERENCES

[1] Saqib Saeed , Sarah A . Suayyid , ManaI s. AI-Ghamdi, Hayfa AI-Muhaisen and Abdullah M.Almuhaideb , ; “A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience”; 19 August 2023.

[2] Megan A. Brown†1,2, Andrew Gruen3, Gabe Maldoff4, Solomon Messing2, Zeve Sanderson2,5, and Michael Zimmer6; “WebScraping for Research: Legal, Ethical, Institutional, and Scientific Considerations ” ; November 1, 2024.

[3] Mustapha Shehu1, Mustafa Man 2, Wan Aezwani Wan Abu Bakar 3, Mohd Kamir Yusof 4, and Ily Amalina Ahmad Sabri ; “A Demystified Overview of Data Scraping ”ISSN: 2563-4429.