C3iHub

POWERED BY

Team name

CybHat

UNCOVERING THREAT INTELLIGENCE FROM CYBERSECURITY REPORTS

Team Members

TEAM LEADER
Sahil Sudhir Dehadray

Djatassiba Yaovi Joel

Abhinav Galagali

HACK IITK
CYBERSECURITY CHALLENGE HACKATHON 2024

Table of content

Introduction

The increasing velocity and sophistication of cyber threats necessitate automated solutions for extracting actionable intelligence from unstructured data sources, such as security reports.

Manual analysis struggles to keep pace with the volume and complexity of cyber threat data

Introduction

Our NLP based solution automates threat intelligence extraction, rapidly identifying IoCs, TTPs, malware references, threat actors, and targeted entities from raw text using regex patterns, SpaCy NER, transformer-based models, and MITRE ATT&CK mappings.

A web-based dashboard enhances user experience with flexible sorting, filtering, and download options.

Problem Statement

1. Unstructured Threat Reports
2. Manual Extraction Challenges
3. Need for Automation
4. Lack of Standardization
5. External Data Enrichment
6. Scalability Issues

Methodology

Threat Intelligence

IOCs

Identifications of critical Indicators of Compromise

- IP Addresses
- MAC Addresses
- Domains
- URLs
- File Hashes (MD5, SHA-1, SHA-256)
- Email Addresses
- Registry Keys
- File Paths
- GUIDs
- Filenames

TTPs

MITRE ATT&CK framework tactics

Identification of Tactics, Techniques, and Procedures

- Reconnaissance
- Resource Development
- Initial Access
- Execution
- Persistence
- Privilege Escalation
- Defense Evasion
- Credential Access
- Discovery
- Lateral Movement
- Collection
- Command and Control
- Exfiltration
- Impact

Malware

Detection of malware names

- Matches known malware name
- Scans report text
- Uses predefined lists
- Queries as Name, hashes, tags

Actors

Recognition of threat actors

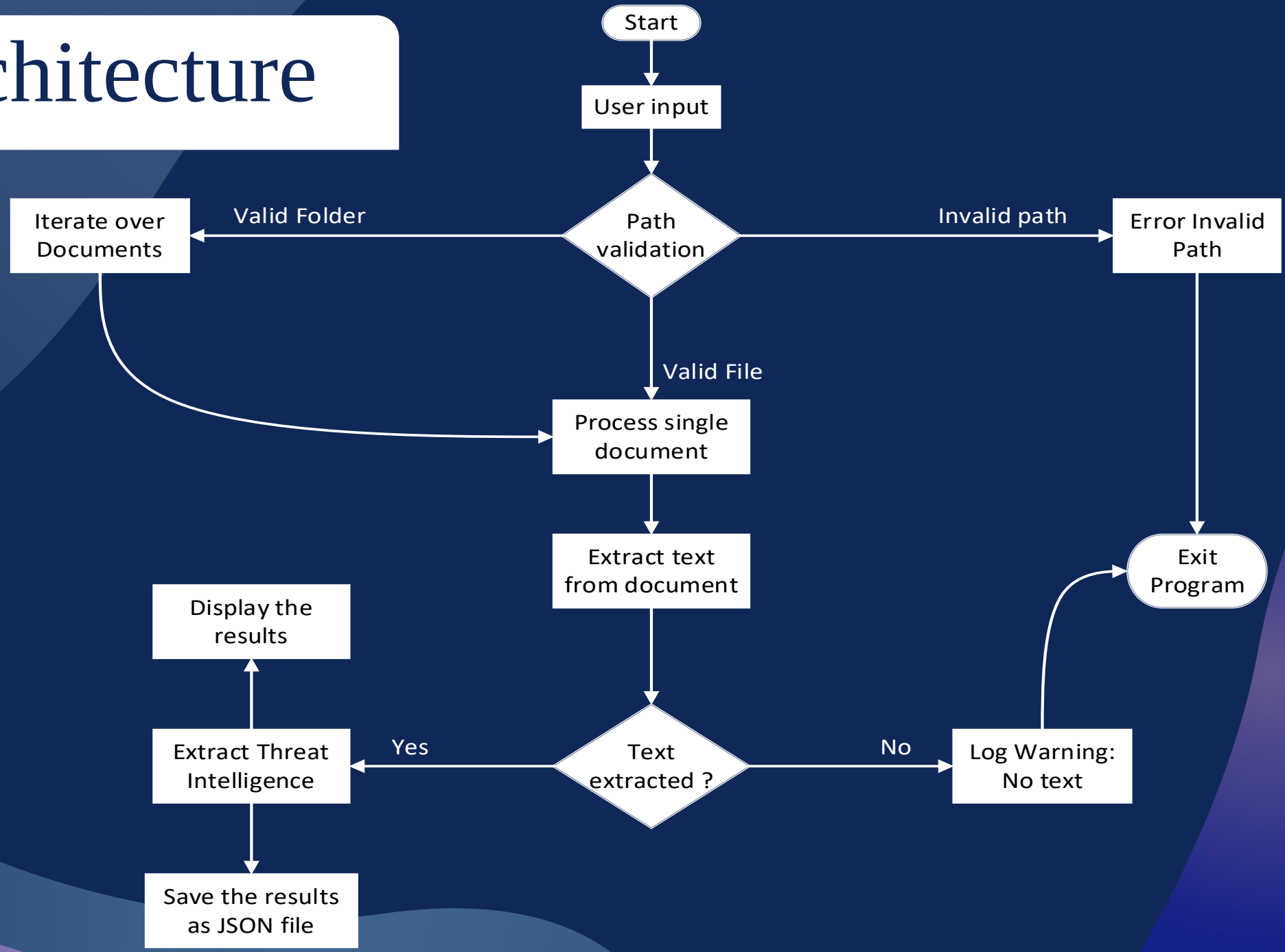
- Uses NER for proper nouns
- Matches known actor names
- Scans report text

Targets

Finding of targeted entities

- Finds targeted entities
- Uses NER for industries/orgs
- Identifies mentioned entities.

Architecture



Technical Requirements

Dependencies

Python 3.7+

spaCy

Werkzeug

PyMuPDF

docx2txt

Torch

requests (APIs)

Transformers

Input Requirements

Machine-readable documents (OCR for scanned files)

Avoid images/tables

Technical Requirements

Key Technologies

HTML

JavaScript

CSS

Python

Flask

Machine Learning (Transformers)

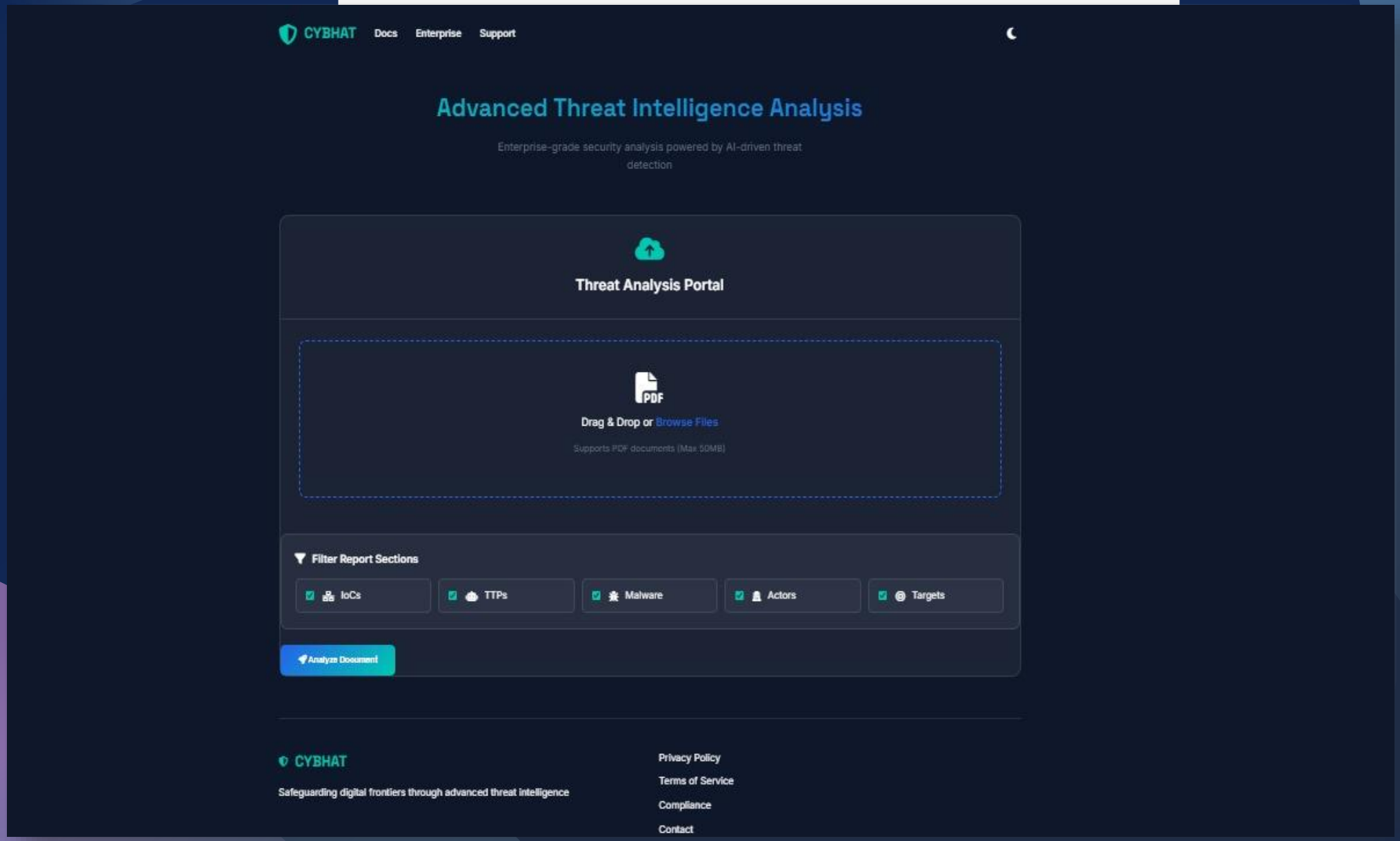
NLP (spaCy)

APIs

Regex

MITRE ATT&CK Framework

Demonstration



Challenges

MITRE ATT&CK Framework Mapping

Researched MITRE ATT&CK tactics techniques and created a manual mapping dictionary

Reducing Errors in Entity Extraction

Combined regex precision with Transformer's NER model flexibility for robust extraction

Differentiating Domains from File Paths

Designed distinct regex rules for domains and file path

Contextual validation to check paths prefixed with drive letters like C:\

Malware Metadata Enrichment

API calls to VirusTotal for malware metadata extraction

Future Scope

AI Assistant Chat Bot

Cloud hosting

Improvement

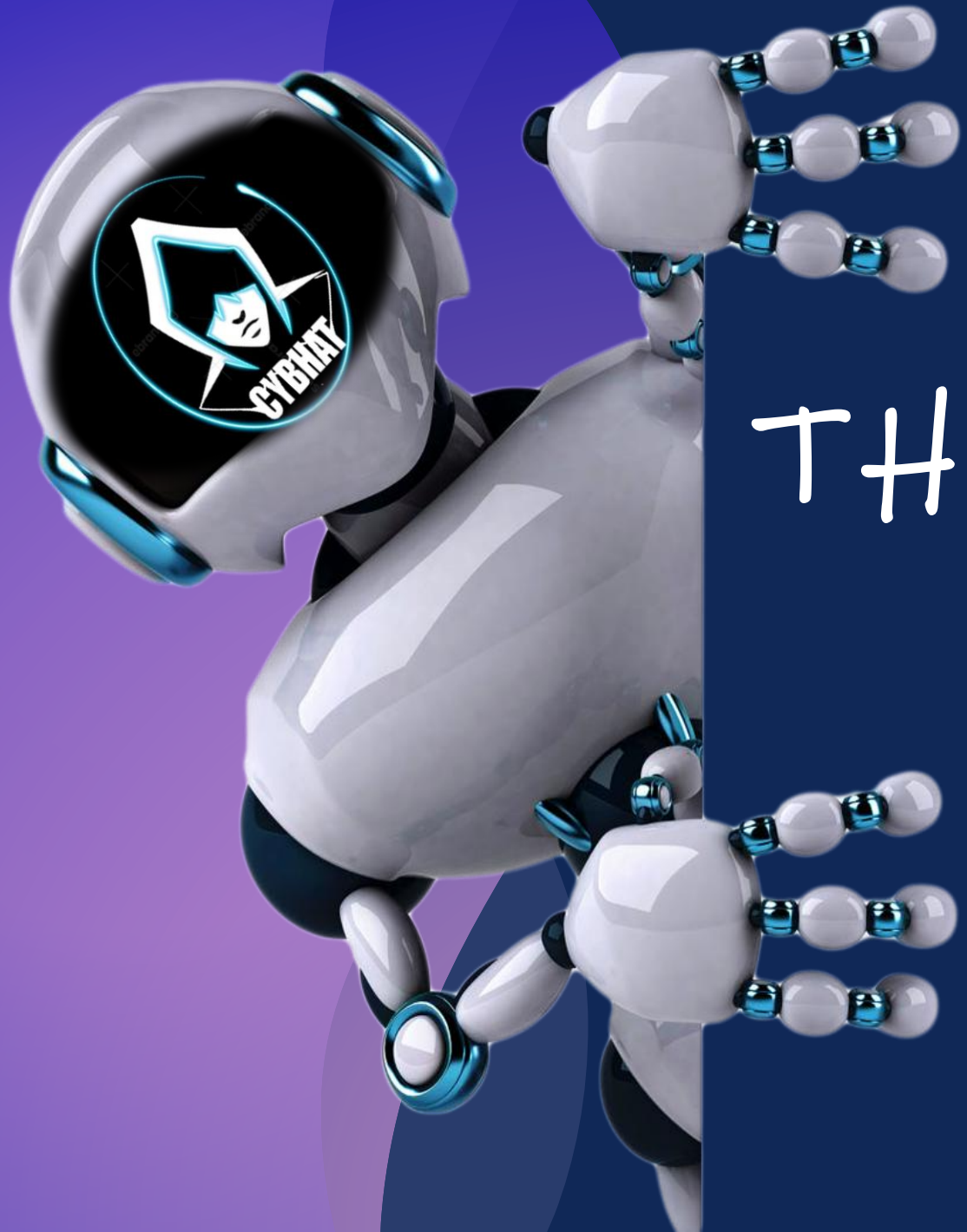
OCR & Image Analyzer Support

Multiple Document Formats

Conclusion

By automating the extraction of threat intelligence with NLP, Machine Learning and MITRE ATT&CK alignment, we bridge the gap between unstructured reports and actionable insight enabling faster, more accurate threat analysis.

As cyber threats escalate, integrating AI models and API-driven enrichment will further refine this process, transforming reactive defenses into proactive, data-driven strategies.



THANK YOU

Q&A Session