

Threat Intelligence Extraction Framework: Technical Report

1. Methodology

This framework automates the extraction of structured threat intelligence from unstructured PDF reports through a multi-stage pipeline:

1.1 Text Extraction

- **Input:** Raw PDF threat reports (e.g., incident analyses, APT campaign summaries).
- **Tool:** `pdfplumber` extracts text while preserving layout context, outperforming alternatives like `PyPDF2` in handling complex PDFs.

1.2 Indicator of Compromise (IoC) Extraction

- **Regex Patterns:** Domain-specific patterns identify:
 - Network artifacts (IPs, domains)
 - Host artifacts (file hashes)
 - Email addresses
- **Deduplication:** Uses Python `set` operations to eliminate duplicates.

1.3 Entity Recognition

- **spaCy NLP Pipeline:** Identifies:
 - **Threat Actors:** Tagged via `ORG/PERSON` entities.
 - **Targets:** Tagged via `GPE` (geopolitical)/`NORP` (nationalities/religions) entities.

1.4 TTP Mapping

- **MITRE ATT&CK Framework:**
 - **Tactics:** Mapped via keywords (e.g., "lateral movement" → `TA0008`).
 - **Techniques:** Sub-technique granularity (e.g., "scheduled task" → `T1053.005`).

1.5 Malware Enrichment

- **VirusTotal Integration:** Fetches:
 - File metadata (SSDeep/TLSH hashes)
 - Behavioral tags (e.g., "ransomware")
 - Threat scores via `last_analysis_stats`.

1.6 Reporting

- **Dynamic Filtering:** Users select fields (e.g., `--fields iocs malware`) via CLI.
 - **JSON Output:** Standardized format for integration with SIEMs (e.g., Splunk, Elasticsearch).
-

2. Key Contributions

2.1 Automation of Manual Processes

- Reduces time-to-analysis from hours (manual review) to seconds.
- Standardizes unstructured data into structured, actionable intelligence.

2.2 Multi-Source Intelligence Fusion

- **Local Analysis:** Regex + NLP for rapid extraction.
- **External Enrichment:** VirusTotal API adds global threat context.

2.3 Precision-Tuned Entity Recognition

- Focused NER categories minimize noise compared to general-purpose models.
 - Example: Filters out non-relevant PERSON entities (e.g., researchers) via post-processing.

2.4 MITRE ATT&CK Operationalization

- Translates prose descriptions to actionable TTPs for:
 - Threat hunting (e.g., hunting for T1059.001 in logs).
 - Incident response playbook development.
-

3. Novelty

3.1 Hybrid Extraction Approach

- **Regex + NLP Synergy:** Combines rule-based precision (IoC regex) with contextual understanding (NER).
 - Solves limitations of pure regex/NLP approaches (e.g., missing novel TTPs).

3.2 Dynamic Filtering

- **User-Driven Output:** Analysts specify fields (e.g., `ioCs ttps`) to align with investigation needs.
- Contrasts with monolithic tools that output fixed data.

3.3 Lightweight Integration

- **Minimal Dependencies:** Uses lightweight libraries (e.g., `spacy` vs. heavyweight LLMs).
 - **API-Agnostic Design:** VirusTotal can be replaced with ANY.RUN or Hybrid Analysis with minimal code changes.
-

4. Technology Justification

Technology	Rationale
pdfplumber	Superior text extraction from multi-column/multi-table PDFs vs. PyPDF2.
spacy	Fast, accurate NER with pretrained models (~10x faster than NLTK).
vt (VirusTotal API)	Industry-standard malware database with 700k+ daily contributors.

Technology	Rationale
argparse	Native CLI support for seamless integration into analyst workflows.
re (Regex)	High-performance pattern matching for time-sensitive IOC extraction.

5. Impact & Applications

- **SOC Automation:** Feed JSON reports into SOAR platforms for automated alerting.
 - **Threat Research:** Accelerates correlation of campaigns with MITRE TTPs.
 - **Red Teaming:** Generates adversary emulation plans from historical reports.
-

6. Conclusion

This framework bridges the gap between unstructured threat reports and machine-readable intelligence. By combining NLP, regex, and external APIs, it enables analysts to focus on high-value tasks (e.g., hypothesis testing) rather than data wrangling. Its modular design ensures adaptability to evolving threats and integration with modern security stacks.