Open-source intelligence (OSINT) is the collection and analysis of data gathered from open sources to produce actionable intelligence.

Technical Cyber Threat Intelligence (CTI) to configure detection systems:
- Indicators to put into SIEMs
- Domains to block in name-servers or proxies
- Execution patterns to block in EDRs

But also “soft” CTI:
- News about threat actors
- New (features of) security products
- News about breaches, incidents, campaigns
- News about vulnerabilities, patches, mitigations, counter-measures, exploitation, post-exploitation, …
- Policy news: political/diplomatic initiatives, new EU policy documents, GDPR-related lawsuits
- Updates on security standards (ISO, BSI, ANSI, CIS, OWASP, …)
- Mergers, acquisitions, failures, … or other company news
WHAT IS OSINT GOOD FOR?

- Gather public information on potential security threats, vulnerabilities, trends, attacker TTPs, new risks etc. to maintain situational awareness and take early counter actions.
- Input for products
  - Advisories
  - Summaries (daily, weekly)
  - Situational reports, white papers, fact-sheets
- Awareness / Preparedness
  - Consulting / Answering calls for help
  - Media inquiries
  - “Boss/CEO/Politician asking questions”
  - Trigger for proactive activities
- Challenge: Number of OSINT sources is high and the number of news items massive
  - Grasp quickly what’s relevant and omit the rest
  - Filter repetitive content
  - The workflow is actually pretty similar to a journalists work
TARANIS AI

• Based on *taranis3* and **taranis-ng**
  • Great tools to ingest raw unstructured data from various sources
  • Use human knowledge to identify relevant information
• Preserves the “taranis workflow” many CERTs are used to
  • Gather -> Assess -> Analyze -> Publish
• Introduces natural language processing (NLP) capabilities
  • Extraction of relevant **named entities**
  • **Clustering** of related **news items**
  • **Summaries** of “story clusters”
  • **Recommendations** of news items
  • Support for creating **OSINT products** (“reports”)

* https://github.com/NCSC-NL/taranis3
** https://github.com/SK-CERT/Taranis-NG

29/05/2024
User Story 1: What are the ‘hot topics’ of the last 24 hours?

User Story 2: What do we know about a specific entity? (e.g., a vulnerability, malware, company, product, person, location etc.)

User Story 3: How can I find more related news items after reading this interesting article?

User Story 4: Which news items are pertinent to my mission?

User Story 5: How can I efficiently sum up my findings for my commander or operators of military IT services?
NOVEL FEATURES AND DEVELOPMENTS

Summary Creation

Advanced Search

Relevance Ranking

NER

Story Clustering

29/05/2024
Named-entity recognition (NER) seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, etc.

• Mix of pre-defined lists (e.g., countries), custom regex (e.g., CVEs), and trained language model on German/English standard text

• Custom extensions to recognize IT products, vendors, APT groups etc.

• Future Extension:
  • Additional use of domain-specific word lists
FEATURES (2/5): ADVANCED SEARCH & FILTERING

- Tags from NER can be used to filter and cluster items belonging to the same topic
- Additional free-text search
- Further filter and sort parameters
  - read/unread, relevance score etc.
  - Sources and source groups
  - Time spans
- Future extension:
  - Collaborative search and filtering
  - Sharing of filters
FEATURES (3/5): RELEVANCE RANKING

- Relevance Ranking helps to identify “interesting items” (aka stories) based on general importance and personal preferences
  - Upvotes and downvotes from collaborators
  - Shared articles
  - Related news items
- Future Extensions:
  - Feedback Loop: Learn properties of often up-/downvotes items.
    - What properties do “good” articles have in common?
    - Which sources deliver such items?
**FEATURES (4/5): SUMMARY & REPORT CREATION**

- Summaries help to condense lengthy texts to their essential parts for quick decisions on their relevance
  - Summary of lengthy news items to quickly grasp its content
  - Summary of stories based on its collection of articles
- **Future Extensions**
  - Summary of Sharing Sets for reports
    - AI-assisted Pre-filling of report fields
  - Tuning of summaries regarding appropriate length, wording and content
• News Items are usually delivered not just by one, but multiple sources at approximately the same time with mostly similar content

• Cluster items and create “meta item” that summarizes important content (“story”)
  • Decrease human effort needed to ingest all news items!
  • Visualize development of a story over time

• Show development of “hot topics” over time
TECHNOLOGY

• Gathering from a multitude of OSINT sources via RSS, from the Web, e-Mail, and some APIs for common issue tracker

• Asynchronous Pub-Sub architecture with workers for flexible extensibility in terms of AI and NLP
  - Workers process items “best effort” in predefined order, e.g., extract IoCs, NER, storyclustering, summary creation etc.

• Resulting products are text file, pdf, html and pushed out via Mail, or to a MISP Server via API

• 100% open source and free to use (EUPL license)
  - Please contribute! (issues, pull request, …)
### LESSONS LEARNED FROM FIRST PILOTS

#### The need for high-quality gathering
- “Garbage in – garbage out”
- Daily digests
- **Polluting texts** (menus, footers…)
- Various formats, changing layouts
- RSS v.s. HTML (in terms of tables, figures, etc.)
- Near duplicates & updated articles

#### The need for adaptable and flexible workflows
- Potentially **complex workflows** across different roles/departments
- **Flexible use cases**
- “generic” features, most important **flags** (has been read, analyzed, reported, escalated, …)
- Preserve the „taranis workflow“, but be adaptable to different organizational structures

#### The need for manually corrected AI-produced results
- AI can tremendously relieve human analysts from tiring tasks
- But **human element remains indispensable**
- Correction of AI-produced results is valuable **feedback to tune algorithms** (revising tags, removing items from story, splitting cluster…)
- Enhances the effectiveness of AI

#### The need for carefully tuned recommender systems
- Even with regular feedback through humans, **learned models remain intransparent**
- Importance of **explainability** – why was an item categorized in a certain way?
- Detect and **mitigate biases and concept drifts**
- Re-evaluation of past decisions is tricky (e.g., dismiss this story for good)
## SCIENTIFIC CHALLENGES & NEW FEATURES AHEAD

- **improve self-asset management** for alignment and matching of findings to own assets (prioritize/rank news items)
- Taking further **context** into account, e.g., mission-specific information, to **provide targeted and easy to comprehend CTI** for efficient & informed decision making in tense and stressful situations
- **use of LLMs to formulate situational awareness reports** for specific stakeholders in their own language, e.g. present information about vulnerabilities, major incidents, new risks for specific roles
- **automatic notification of emerging threats**, e.g., because of dynamic growth of a story on a specific topic, **sentiment analysis** of social media posts that mention entities in a certain context
- **improve the handling of classified information**, e.g., stories from OSINT that are being enriched with closed source information can only be partially disseminated and via pre-defined channels
- **improve sharing capabilities** across instances of taranis-AI (currently analysts work all on one instance), standardized interfaces to common solutions in the domain, e.g., MISP
FINANCIAL SUPPORT AND COLLABORATIONS

• The work in this presentation was mainly funded by the Connecting Europe Facility (CEF) program in course of the project AWAKE (2020- AT-IA-0254).

• The work in this paper has additionally received funding from the European Union - European Defence Fund under GA no. 101121418 (EUCINF) and 101121403 (NEWSROOM). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

• Active collaboration with CERT.at, Austrian Ministry of Interior and Austrian MoD
  • Early adoption of taranis-ai by analysts

• Exchange with SK CERT, the developers of taranis-ng

• Open Source project on Github
  • Opportunity to contribute: https://taranis.ai/
THANK YOU!

Please contact:
Florian Skopik
florian.skopik@ait.ac.at

May, 30th 2024