International Journal of Information Security
https://doi.org/10.1007/s10207-024-00860-w

REGULAR CONTRIBUTION

Integrating AI-driven threat intelligence and forecasting in the cyber security exercise content generation lifecycle

Alexandros Zacharis[1] · **Vasilios Katos**[2] · **Constantinos Patsakis**[3,4]

Abstract
The escalating complexity and impact of cyber threats require organisations to rehearse responses to cyber-attacks by routinely conducting cyber security exercises. However, the effectiveness of these exercises is limited by the exercise planners' ability to replicate real-world scenarios in a timely manner that is, most importantly, tailored to the training audience and sector impacted. To address this issue, we propose the integration of AI-driven sectorial threat intelligence and forecasting to identify emerging and relevant threats and anticipate their impact in different industries. By incorporating such automated analysis and forecasting into the design of cyber security exercises, organisations can simulate real-world scenarios more accurately and assess their ability to respond to emerging threats. Fundamentally, our approach enhances the effectiveness of cyber security exercises by tailoring the scenarios to reflect the threats that are more relevant and imminent to the sector of the targeted organisation, thereby enhancing its preparedness for cyber attacks. To assess the efficacy of our forecasting methodology, we conducted a survey with domain experts and report their feedback and evaluation of the proposed methodology.

Keywords Cyber security exercise scenarios · Machine learning · Threat intelligence · Threat forecasting

1 Introduction

Cyber Security Exercises (CSE) have gained prominence in the training landscape, providing hands-on experience to personnel across various industries. A CSE, as described in the ISO Guidelines for Exercises [31], is "*a process to train for, assess, practice, and improve performance in an organisation*". However, the effectiveness of these exercises is often hindered by the inability to replicate real-world scenarios in a timely manner relevant to the sector methodology. Our research aims to identify and bridge the gap in context-relevant CSE scenarios generated by sectorial cyber security experts and seasoned exercise planners. To address this challenge, this work proposes integrating AI-driven sectorial tag coverage techniques into the CSE scenario generation process. Although threat intelligence and forecasting have been used for various cyber security tasks, including CSE scenario design, our fully automated AI-driven methodology is the first of its kind to incorporate forecasting and scenario generation in a single flow. Using our approach, organisations can generate sector-specific threat intelligence that improves the realism and relevance of future CSE scenarios, especially when integrated into exercise generation frameworks such as the AI-assisted Cyber Exercise Framework (AiCEF) [66].

The research questions we aim to address it this work are:

- Can we produce reliable sectorial threat landscapes using AI models and methodologies?
- How much can experts and nonexperts developing CSE scenarios benefit from our methodology?

To address these questions, the following objectives have been set for this work and can be summarised as follows:

✉ Constantinos Patsakis
kpatsak@unipi.gr

Alexandros Zacharis
alexandros.zacharis@enisa.europa.eu

Vasilios Katos
vkatos@bournemouth.ac.uk

1 European Union Agency for Cybersecurity (ENISA), Chalandri, Greece

2 Bournemouth University, Poole, UK

3 Department of Informatics, University of Piraeus, Karaoli & Dimitriou 80, 18534 Piraeus, Greece

4 Athena Research Center, Marousi, Greece

Published online: 10 May 2024

1. Develop an AI-driven threat intelligence and forecasting methodology that accurately identifies current and emerging trends with their potential impact in different sectors from any given dataset and incorporates them into a CSE scenario. To achieve this objective, we evaluate several existing forecasting algorithms using quantifiable criteria to identify those that provide accurate predictions for a given dataset.
2. Evaluate the effectiveness of integrating AI-driven sectorial threat intelligence and forecasting into the exercise generation process by comparing the realism and relevance of CSEs created using our methodology versus the traditional human experts' approach. Additional effort was devoted to implementing our methodology into a fully automated prompt generation methodology that non-expert exercise planners can easily use. To achieve this objective, we have conducted surveys based on a case study to measure the perceived relevance of CSE with AI-driven threat intelligence and forecasting compared to traditional exercise generation. No other AI-driven exercise generation framework exists against which we can compare our methodology. For this reason, we compared our methodology against human-based approaches by tasking experts in this field. We have then collected quantitative results that help us compare our methodology against the human expert's methods used in identifying current and emerging threats and compare the results against those of renounced threat landscapes used as a baseline.

Our analysis, coupled with valuable insights from field experts, has yielded tangible results affirming that the integration of our proposed methodology into the Cyber Security Exercise generation process significantly improves the realism and relevance of CSE scenarios. This enhancement can drastically improve organisations' ability to confront real-world cyber threats effectively.

The structure of the remainder of this work is as follows. In Sect. 2, we delineate the current state-of-the-art about on AI-driven techniques to enhance cyber security exercises and improve organisations' preparedness for emerging threats, focusing on machine learning techniques used for threat forecasting. Our methodology is presented in Sect. 3, where we deep dive into its implementation, focusing on data collection, data and trend analysis, topic modelling and extraction. We then compare different forecasting and trend prediction models and evaluate them for use in the following steps. All previous steps are incorporated into our AI-driven prompt generation (AiDPG) tool, which can be used to create tailored CSE scenarios. We present our evaluation results in Sect. 4, using a case study focusing on the energy sector and involving both sectorial and exercise experts. We conclude this article by presenting the lessons learned and future work in this area of interest in Sect. 5.

2 Related work

The following paragraphs provide a brief overview of the related work. First, we discuss the relevant forecasting models, focusing only on the machine learning models we used in our study. Then, we discuss the trends in generating CSEs and the use of forecasting methods.

2.1 Forecasting models

The main forecasting models that we consider in the scope of this work, along with their limitations, are the following.
SARIMAX (Seasonal Auto Regressive Integrated Moving Average with Exogenous Variables) [9] is a popular statistical model for time series forecasting. It extends the ARIMA (AutoRegressive Integrated Moving Average) [9] model by incorporating seasonal patterns and exogenous variables. Unlike ARIMA, which effectively analyses and predicts time series data with stationary or non-stationary properties, SARIMAX captures the relationships between past observations, forecast errors, and external factors to provide more accurate forecasts. However, SARIMAX models assume linearity and stationarity in the data, which may not always be true in real-world scenarios. SARIMAX models do not handle missing data or outliers well, which requires additional preprocessing steps. Interpreting the coefficients and understanding the impact of exogenous variables can be complex.
Exponential Smoothing (ES) [28, 30] is a widely used time series forecasting technique that assigns exponentially decreasing weights to past observations. It is based on the assumption that recent observations have greater significance in predicting future values. Yet, ES may struggle to capture complex non-linear patterns in the data.
Simple Exponential Smoothing (SES) [29] is a specific form of exponential smoothing that only considers the current observation and a single smoothing parameter (alpha). It provides a straightforward method for forecasting by exponentially weighting past observations and updating the forecast based on the weighted average.
Nonetheless, SES assumes a constant level and does not capture trends or seasonality, making it less suitable for data with such patterns. Moreover, it may struggle with data that have irregular or changing patterns over time and is more suitable for short-term forecasting rather than long-term predictions.
Prophet [58] is a forecasting framework developed by Facebook's Core Data Science team. It combines the components of additive regression models with flexible seasonality estimation. Prophet incorporates trend components, seasonal

effects, and additional regressors to generate forecasts. It also effectively handles outliers and missing data points. However, Prophet assumes that trends and seasonality are additive and follow specific patterns, which may not always hold in diverse datasets. The automatic change-point detection in Prophet might not always accurately identify complex or abrupt changes.

Long Short-Term Memory (LSTM) [20, 24] is a type of **recurrent neural network (RNN)** architecture commonly used to model and forecast time series data. LSTMs are designed to overcome the limitations of traditional RNNs in capturing long-term dependencies and handling vanishing or exploding gradients. In LSTM models, the network learns to selectively retain or forget information from past observations, allowing it to capture and retain important patterns over longer sequences.

On the other hand, training LSTM models requires more data to capture long-term dependencies and avoid over-fitting effectively. LSTM models can be computationally intensive. Hyper-parameter tuning for LSTM models can be challenging, as the optimal architecture and parameter settings may vary depending on the specific dataset and problem.

2.2 Cyber security exercise scenarios and forecasting

The integration of AI-driven techniques into CSEs has gained significant traction in recent years as organisations seek to enhance their preparedness for emerging threats, enhancing traditional CSE scenario-building methodologies. AI-assisted cyber exercise content generation framework (AiCEF) [66] harnesses the power of Named Entity Recognition (NER) to extract tags based on a Cyber Exercise Scenario Ontology (CESO) through its Machine Learning to Cyber Exercise Scenario Ontology (MLCESO) conversion module. Successful CSEs hinge on the effectiveness of their scenarios, which must simulate real-world situations that resonate with participants while maintaining a level of realism that fosters engagement and learning. Realistic scenarios, as conceptualised in traditional exercise scenario-building methodologies [5, 17], serve as powerful tools for predicting future events, incorporating relevant issues, interactions, and potential consequences. Examples of realistic scenarios modelled from past cyber conflicts can be found in [67] partially foreseeing future cyber events. These scenarios, according to Green and Zafar [23] and Granåsen et al. [22], lead to constructive training experiences that enhance participant understanding and preparedness.

More concretely, an exercise scenario is a sequential narrative account of a hypothetical incident that catalyses the exercise and intends to introduce situations that will inspire responses and thus allow demonstration of the exercise objectives [51]. ENISA acknowledges the benefits of sectorial exercises scenario in their National Exercise - Good Practice Guide [12], an approach recognised by national cyber security authorities like ANSII in their guide on organising cyber crisis management exercises [3]. In the context of CSEs, a scenario defines the training environment that will lead participants towards fulfilling the exercise objectives set [33]. The cyber security problem described in the scenario itself portrays a structured representation named Master Scenario Events List (MSEL), which serves as the script to execute an exercise [51].

To generate a useful MSEL, we need to successfully detect and classify cyber attacks relevant to the needs of the trainee, as emphasised in [36, 47, 49]. Interestingly, applying advanced machine learning techniques to predict future cyberattack trends has gained significant momentum in cyber security. Long Short-Term Memory (LSTM) networks constitute a specific type of recurrent neural network that has been particularly noted for their ability to learn long-term dependencies and patterns in time-series data, which is characteristic of cyber attack vectors. For instance, [34] demonstrated the efficacy of LSTMs in classifying intrusion detection patterns, emphasising their potential to identify sophisticated cyber threats by analysing historical data patterns. Similar studies to ours can be considered in the cases of [1, 6, 15, 19] where LSTM networks are proposed as a method to forecast cyber incidents with a focus on malware infections.

Parallel to the application of LSTMs, the Neural Prophet model has emerged as a novel approach for time-series forecasting [58]. Although specific studies focusing on the use of this model in cyber attack predictions are scarce, its robustness in detecting complex patterns is essential to predict cyber incidents, as reported by [40] with promising results. In the realm of more traditional statistical methods, ARIMA models have been a mainstay for time-series forecasting and have found applications in cyber-security. Sarker et al. [50] discuss integrating machine learning approaches like ARIMA in cyber-security, highlighting its effectiveness, especially in scenarios with limited computational resources or where model interpretability is vital. Following the same path, data mining and machine learning techniques can be applied to VSE scenario development to help develop a mitigation plan for future incident trends and cyber security incident timelines as proposed in [26].

Further to analysing existing threats, modelling [64] predicting upcoming cyber incidents [37, 55] and the overall threat trends are equally important. Several studies focus on the short-term prediction of the number or source of attacks to be expected in the following hours or days [2, 7, 10, 11, 21, 25, 27, 37, 39, 41–43, 45, 61–63]. Most of these works make predictions in restricted settings; e.g., against a specific entity or organisation for which historical data are available [39, 43, 63]. Forecasting attack occurrences has been attempted with statistical methods, especially when parametric data distribu-

tions could be assumed [62, 63], as well as by using machine learning models [21] and deep learning [52]. Other methods adopt a Bayesian setting and build event graphs suitable for estimating the conditional probability of an attack following a given chain of events [45]. Such techniques rely on libraries of predefined attack graphs to identify the known attacks that are most likely to occur. Research has also covered zero-day attacks [54, 57] with less impressive results.

Almahmoud et al. [2] introduced a machine learning based proactive approach for long-term cyber-attack prediction, allowing early detection of potential threats. This approach could facilitate cyber security experts in prioritising resources and mitigation measures, including training through CSEs, a type of proactive cyber defence [56].

The existing literature showcases the potential for AI-driven topic extraction and forecasting techniques to be applied in the identification and prioritisation of both current and future threats in the context of CSEs, minimising their impact upon materialisation [7, 16]. Our work is the first to infuse the outputs of fully automated AI-driven threat intelligence and forecasting into the CSE content generation life cycle. Comparative results are provided against the current, traditional CSE scenario development methodologies used, which are a combination of manual and human-driven effort, underlining the potential of our methodology.

3 Proposed methodology

3.1 Generic overview

While in this section, we outline the proposed methodology for analysing a dataset of cyber security incidents using the AiCEF framework, specifically the MLCESO module; it is generic enough to allow its individual use in other domains and applications. Additionally, we propose prompts based on these trends, prioritising the needs per sector through our AI-driven Prompt Generation (AiDPG) methodology. In terms of the threat intelligence collection methodology used, we decided to follow some of the steps proposed by ENISA [13] to create our own sectorial threat intelligence (STI) report with the difference of using AI exclusively as a means of data processing instead of human analysts. In what follows, when referring to sectors, we will refer to those considered in the NIS 2 Directive [14]; however, any other such classification can be used. The outline of our methodology is illustrated in Fig. 1, consisting of five distinct steps.

In what follows, we assume access to a comprehensive and representative dataset of cyber security incidents, in text format, that spans over several years. Moreover, we assume that all incidents are unique to prevent biases in our predictions. Nevertheless, to exemplify, as we discuss in the following paragraphs, we point the reader to a manually curated dataset,

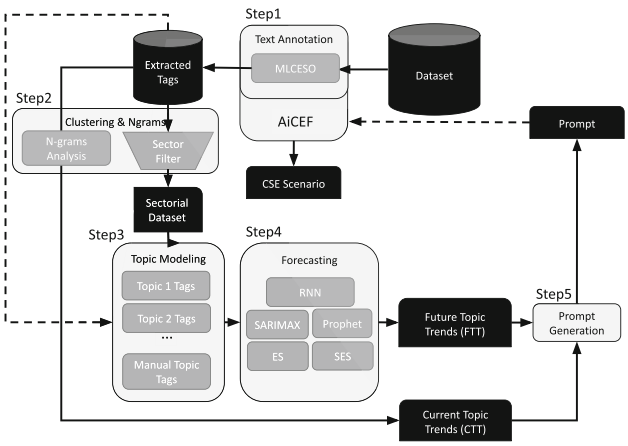

Fig. 1 Proposed methodology overview

and we use it as a reference for showcasing and validating our methodology.

1. **Text annotation using AiCEF** The AiCEF framework is employed, being the only available CSE-specific framework, and we use the MLCESO module for preprocessing (Paragraph 3.2) and text annotation, extracting relevant exercise-related objects, such as attacker type, attack type, and exploited vulnerabilities (Paragraph 3.3).
2. **Clustering and N-Gram Analysis** Incidents are clustered into sectors based on the extracted word tags, to identify similarities and group similar incidents for a more detailed N-gram analysis (Paragraph 3.4).
3. **Topic Modelling** We apply the Latent Dirichlet Allocation (LDA) algorithm for topic modelling, extracting topics within the context of elements that might be involved in a CSE scenario (Paragraph 3.5). Given the increasing use of large language models (LLMs) and their exceptional results in multiple domains [18, 32, 44, 59], we anticipate that soon LLMs will be able to automate these steps of our methodology relatively efficiently.
4. **Forecasting** Occurrence analysis is performed to capture the incident frequency, patterns, and trends over time, extracting insights about the temporal aspects of the data. The latter allows us to identify current and future trends, patterns, anomalies, and emerging trends in cyber security incidents within each sector. Although any single time series analysis algorithm can be used for this step, we decided to put SARIMAX, ES, SES, Prophet, and LSTM to the test and evaluate them using R-squared (R^2) [35], mean absolute error (MAE) [35], mean absolute deviation (MAD) [60] and mean absolute percentage error(MAPE) [4] metrics to identify the best-performing ones for this task (Paragraph 3.6).
5. **Prompt Generation** An exercise prompt is considered a detailed representation of a CSE scenario, containing

all the important cyber security elements that will be later taken into consideration to draft a CSE. Actionable prompts are proposed back to AiCEF via our AiDPG methodology, using the identified trends to address cyber security challenges and mitigate risks but also prioritising training topics based on sector-specific requirements and characteristics (Paragraph 3.7).

3.2 Dataset and data preprocessing

As shown in Table 1a, four incident sources have been identified as input to generate our corpus. These websites contain unique articles previously selected and classified in the Hackmageddon dataset,[1] a reputable webpage that describes cyber security incidents since 2011. For simplicity, in this work, we collected unique incidents from 2019–01 to 2022–12 (Table 1b), which accounts for 4.957 articles. We spanned our collection over four years to adhere to the limitations paused by the time series analysis algorithms to be used in later steps. Note that while there were undoubtedly attacks before 2019, neither the amount of information nor the depth can be comparable to today's figures. The issuance of the Presidential Policy Directive 41 (PPD-41) in 2016 in the USA, titled "United States Cyber Incident Coordination" and in parallel with the introduction of the General Data Protection Regulation (GDPR) in the same year, were the main drivers in pushing organisations both private and public to report cyber security incidents. Therefore, even if significant cyber security incidents existed before 2016, they were not reported, let alone published. As a result, if one simply compares the number of reports before 2019 that GDPR was enforced, one would notice a massive surge in the numbers, as the bulk of incidents were hidden under the carpet. Of course, that would introduce severe biases in our dataset. Therefore, we opt to use a smaller yet more representative dataset.

All relevant articles were collected through automated web scraping. Then, the plain text of each article was processed using Natural Language Processing (NLP) techniques to form the reduced Incidents Corpus (IC). Initially, all texts were converted to the UTF-8 encoding scheme. Using dictionaries and the Textblob library,[2] we corrected spelling and removed special characters. Empty lines, specific stopwords, and punctuation marks were removed using traditional NLP libraries such as NLTK[3] and spaCy. [4] In addition, all HTML and other programming codes, URLs, and paths were removed.

The standard Penn Treebank [38] tokenisation rules were used for sentence tokenisation to tune the incidents' text and facilitate annotation. At the end of this step, a corpus composed of timestamped incidents was formed.

Table 1 Composition of our dataset

(a) dataset quota per source

Source	# Incidents
`bleepingcomputer.com`	2431
`securityaffairs.co`	339
`zdnet.com`	794
`databreaches.net`	1393
Total	**4957**

(b) Dataset yearly distribution

Year	# Incidents
2019	1088
2020	1348
2021	1248
2022	1353
Total	**4957**

3.3 Text annotation

To annotate our dataset, pre-trained named entity recognition models (NER) (Table 2) were used. These models are part of AiCEF's MLSECO module and can annotate text with predefined tags, as described in Table 3. The output of the annotation step was stored in a Knowledge Database (KDb). An example of the information stored in KDb can be seen in Table 4.

3.4 N-gram analysis

By applying NER and extracting important keywords related to the above categories, we have already taken a crucial step in preparing the data for N-gram analysis, a computational technique to extract meaningful patterns and insights from text data. It involves identifying and analysing sequences of N consecutive words (N-grams) that occur frequently within a given corpus. In the context of processing cyber security incidents, N-gram analysis can provide valuable information about the prevalent attack types, techniques, attacker names, malware types, assets, vulnerabilities, and sectors affected by these incidents. Thus, we can leverage this technique to gain a deeper understanding of the relationships and trends in the cyber security incidents that we have in our database.

By performing a top-ten terms analysis (for Attack Types, Techniques, Attackers, Malware Types, Assets, Vulnerabilities, and Sectors) over four years' worth of data in KDb, we were able to visualise the results after clustering similar notion phrases, where deemed necessary. Furthermore, we can uncover relationships between different categories. For

[1] https://www.hackmageddon.com

[2] https://github.com/sloria/TextBlob

[3] https://www.nltk.org/

[4] https://spacy.io/

Table 2 AI models' scores

Category	Tag	Presicion	Recall	F1
Attacker	ATTACKER_TYPE	100.00	83.33	90.11
	ATTACKER_NAME	95.29	87.10	91.01
	ATTACKER_ORIGIN	Used Native Spacy LOC tag (no training)		
Attack	MALWARE_TYPE	80.56	76.32	78.38
	MALWARE_NAME	95.29	87.10	91.01
	ATTACK_TYPE (TECHNIQUE)	88.60	87.07	87.83
	VULNERABILITY	87.50	84.00	85.71
Victim	SECTOR	85.84	84.07	84.95
	ASSETS	87.02	89.06	88.03
	TECHNOLOGY	87.60	89.93	88.70

Table 3 Annotation tags per category

Category	Tag	Link to CESO & STIX 2.1
Attacker	ATTACKER_TYPE	Threat Actor Attribute
	ATTACKER_NAME	Threat Actor Attribute, Identity
	ATTACKER_ORIGIN	Location
Attack	MALWARE_TYPE	Malware Attribute
	MALWARE_NAME	Malware Attribute
	ATTACK_TYPE (TECHNIQUE)	Attack Pattern
	VULNERABILITY	Vulnerability
Victim	SECTOR	Identity Attribute, Scenario
	ASSETS	Threat Actor Attribute
	TECHNOLOGY	Tool

instance, frequent co-occurrence of N-grams like "APT3" and "PlugX malware" may indicate the involvement of specific advanced persistent threat groups and the utilisation of sophisticated malware in cyber attacks.

To showcase an example of threat intelligence extracted for the given dataset, we provide the output of our methodology for the sector tag. After analysing the 4957 cyber security incidents, the following distinct clusters or words representing at least seven heavily impacted sectors were formulated:

- **Commerce:** COMMERCE, COMPANY, COMPANIES, BUSINESS, BUSINESSES, RETAIL SECTOR, ENTERPRISE, ENTERPRISES, ORGANIZATIONS, ORGS, RETAIL, FIRMS, FIRM
- **ICT&Technology:** E-COMMERCE, GAMING, INFRASTRUCTURE, ITECH, ELECTRONICS, IT SERVICES, INTERNET SERVICE PROVIDERS, MARKETPLACE, ONLINE SERVICES, INFORMATION TECHNOLOGY, CLOUD HOSTING, SERVICE PROVIDER, HOSTING, TECHNOLOGY, TECH, CLOUD STORAGE, CLOUD STORAGE SERVICES, NETWORK OPERATOR, IT, TELECOMS, TELCO, TELECOMMUNICATIONS PROVIDER, TELECOMMUNICATIONS, TELECOM, CLOUD SERVICES
- **Finance:** BANKS, STOCK EXCHANGE, BANK, FINANCE, FINANCIAL, BANKING SECTOR
- **Healthcare:** HEALTHCARE, CLINICS, CLINIC, HOSPITAL, MEDICAL PROVIDERS, HEALTHCARE SYSTEM, HEALTH, MEDICAL CENTER, MEDICAL, PHARMACEUTICAL
- **Education:** EDUCATION, SCHOOLS, SCHOOL, UNIVERSITY, UNIVERSITIES, COLLEGE, EDUCATION SECTOR, EDUCATIONAL, HIGHER EDUCATION
- **Government:** GOVERNMENT, PARLIAMENT, MINISTRY, MINISTRIES, GOVT, GOVTS, GOVERNMENTS, GOVERNMENT AGENCIES, POLITICAL, MUNICIPALITIES, ELECTION, CITY, CITIES, FEDERAL AGENCIES

Further to this, smaller clusters of words representing less impacted sectors also emerged, as seen below:

- **Energy:** ENERGY, ENERGY SECTOR, NUCLEAR, OIL, GAS, POWER, PETROL, POWER SUPPLY
- **Defence:** DEFENCE, DEFENSE, LAW ENFORCEMENT, LAW ENFORCEMENT AGENCIES, LAW ENFORCEMENT AGENCY, POLICE, MILITARY

Table 4 Annotation example using AiCEF (MLCESO)

Field	Value
id	2246
Name	202203-2021-0153-58be06f8-f18c-4e1b-a243-9ecb13442636
URL	https://www.bleepingcomputer.com/news/security/bluenoroff-hackers-steal-crypto-using-fake-metamask-extension/
Text	BlueNoroff hackers steal crypto using fake MetaMask extension..
Month	1
Year	2022
Maturity	150
MetaTags	LINUX, SCRIPTS, HACKS, POWERSHELL, CRYPTOCURRENCY, THREAT ACTOR, PASSWORD, MALWARE, EXCEL, WORD, USER CREDENTIALS, CREDENTIALS, POWERSHELL, CHROME EXTENSION, WINDOWS, HACKERS, FILES, ATTACKERS, NORTH KOREA, BROWSER, BROWSER EXTENSIONS, CVE-2017-0199, DATA BREACH, TEMPLATE INJECTION, WALLET, FINANCIAL, GROUP, VULNERABILITY, BLUENOROFF, THREAT ACTORS, DATA
Attacker_Name	BLUENOROFF
Attacker_Type	HACKERS, GROUP, THREAT ACTOR
Threat_Actor_Type	crime-syndicate
Attack_Type	DATA BREACH
Technique	CREDENTIALS, POWERSHELL
Vuln	TEMPLATE INJECTION, VULNERABILITY, CVE-2017-0199
Malware_Name	
Malware_Type	SCRIPTS, MALWARE
Attacker_Origin	NORTH KOREA
Sector	FINANCIAL
Assets	FILES, WALLET, USER CREDENTIALS, CRYPTOCURRENCY, PASSWORD, DATA
Technology	WINDOWS, LINUX, EXCEL, WORD, BROWSER, POWERSHELL, BROWSER EXTENSIONS, CHROME EXTENSION

- **Aerospace:** SPACE, AEROSPACE, AVIATION
- **Media:** MEDIA, MEDIA OUTLETS, NEWS AGENCY, NEWSPAPER, RADIO STATION
- **Maritime:** MARITIME, PORT, SHIP, SHIPS
- **Food:** FOOD, FOOD CHAIN

According to the analysis dataset from 2019 to 2022, the ten largest sector clusters affected are represented per yearly occurrence in total volume in Fig. 2. Since some incidents could have impacted more than one sector, they appear in multiple sectors affected.

Following the same process for all other labels: Assets, Vulnerabilities, Attack Types, Attacker Names, Malware Types, and Malware Names, we can derive the respected threat intelligence. This information is then stored in our KDb per sector, providing us with the Sectorial Threat Intelligence (STI) needed to enrich CSE scenarios.

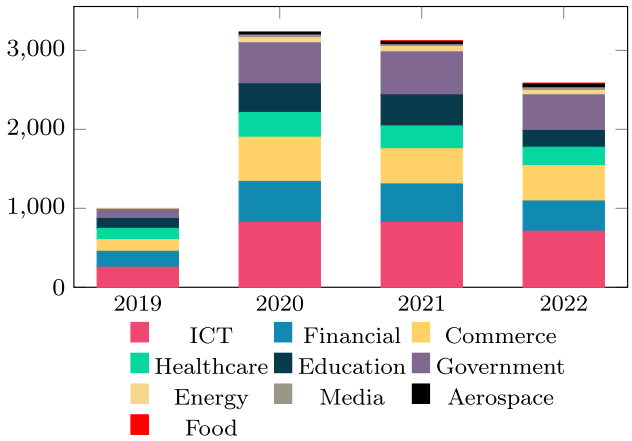

Fig. 2 Top 10 sectors impacted between 2019 and 2022 in volume

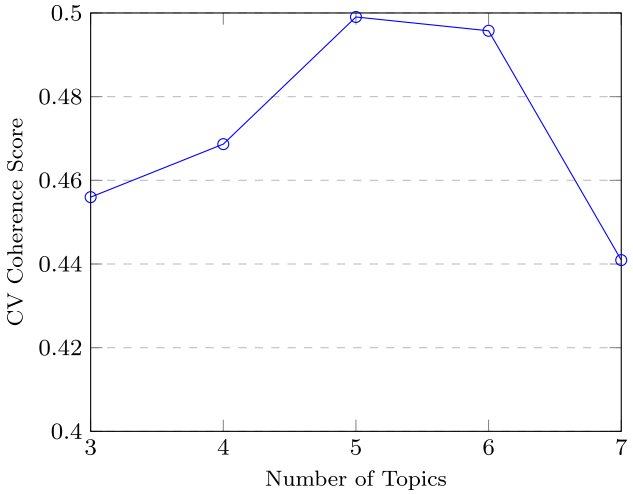

Fig. 3 CV coherence score for different numbers of topics

3.5 Clustering through topic modelling

Topic modelling, specifically key phrase cluster analysis, plays a significant role in understanding and extracting meaningful insights from text corpora. Although any topic modelling algorithm like BERTopic and non-negative matrix factorization (NMF) can be used for this step, we decided to incorporate LDA in our methodology. [8, 46, 53]. LDA is a generative statistical model used to discover topics within a collection of documents. It assumes that each document is a mixture of various topics, and each topic is a distribution over words. LDA scales well with large datasets, making it suitable for analysing extensive collections of documents. LDA is often preferred when dealing with a large text corpus, as NMF can be computationally more expensive. BERTTopic, like NMF requires more computational resources, however, the interpretability of the extracted topics can be challenging, especially compared to methods like LDA [65].

To properly implement the LDA algorithm, the optimal number of topics must be first identified based on a coherence score usually calculated using coherence models such as the CV coherence [48]. The coherence score measures the degree of semantic similarity between words within each topic, and its value can range from zero to one. Comparing the coherence scores obtained for different numbers of topics, the number of topics that yields the highest coherence score is considered the optimal number of topics for the given corpus.

Following the above steps, we conclude that the optimal number of topics to be extracted for the given dataset is five (5), as illustrated in Fig. 3.

Once we trained our LDA model and settled for five topics, we extracted a set of unique words per topic, allowing us to categorise them with specific tags, reported in Table 5

3.6 Forecasting, trend prediction and model evaluation

As already discussed, we tested various time series analysis methodologies, such as SARIMAX, Exponential Smoothing (ES), Simple Exponential Smoothing (SES), Prophet, and Long Short-Term Memory (LSTM) modelling using Recurrent Neural Networks (RNN), following recommendations of previous research findings [2]. Nonetheless, our goal was not to perform an exhaustive comparison of all possible time series analysis methodologies but to assess whether time-series analysis can support our effort of generating relevant cyber security scenarios with a limited incident dataset (of up to four years in this case) of cyber security incidents in hand. To this end, we performed trend prediction using data for three years (2019–2021) of the available dataset to predict the values for the fourth year (2022). Then, to evaluate our models, we focused on four metrics, namely: R-squared (R^2), mean absolute error (MAE), mean absolute deviation (MAD) and mean absolute percentage error(MAPE).

1. R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s) and was used to assess the goodness of fit of our regression models.
2. MAE represents the average absolute difference between the predicted and actual values and provided use with a straightforward measure of prediction accuracy. (Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics were also considered for the same task but were dismissed due to the tasks sensitivity to outliers, where MAE is preferred)
3. MAPE expresses the average percentage difference between predicted and actual values, providing a relative measure of accuracy. MAPE is the most used measure for forecasting error since the variable's units are scaled to percentage units, which makes it easier to understand.
4. MAD is similar to MAE and is used in forecasting to quantify the accuracy of a forecasting model. It represents the average magnitude of forecast errors and helps assess how well the model predicts future values.

As a general guideline, having at least two yearly seasonal data cycles is often recommended to capture seasonal patterns effectively. Assuming that a single seasonal cycle corresponds to one year for monthly data, we would ideally want to have a minimum of 24 observations (two years) to perform a meaningful time series analysis. Theoretically, this allows the models to capture seasonal variations and assess their significance. Clearly, having more observations improves the accuracy and reliability of the analysis and forecasting results. Additionally, the complexity of the time series patterns and the specific goals of the analysis may also impact

Table 5 Topics and Content

Unique words	Tag
HEALTH, RECORD, RECORDS, PHI, PI, PERSONAL, DATA, BIOMETRIC, SSNS, BREACH, DATABASE, ACCOUNT, DETAILS, IDENTIFICATION, FILES	Personal Data Breaches
PHISHING, SPEAR-PHISHING, SPEAR, WALLET, ETHEREUM, SCAM, SIM, SWAPPING, SPOOF, IMPERSONATION, LOGINS, EMAIL, MEDIA, ACCOUNT, LINK	Phishing & social engineering
DDOS, DOS, EXPLOITATION, INFRASTRUCTURE, WINDOWS, LINUX, WEB, INTERNET, DENIAL, SSH, NETWORK, IP, SERVER, ROUTER, CHAIN, SUPPLY, SHELL, INTERNET, PROXY, DNS	Network and System Exploitation
CREDIT, CARD, CARDS, CREDENTIALS, DATA, BANKING, PASSWORD, PAYMENT, DATABASE, SKIMMING, SKIMMER, LOGIN, KEYLOGGING, LINK, COOKIE, DEBIT	Payment Card Fraud & Online Theft
ACCESS, REMOTE, RAT, TROJAN, EXTORTION, RANSOMWARE, TROJANS, VIRUS, SCREEN, VOICE, VIDEO, CLIENT, SCREENSHOTS, ZERO-DAY, O-DAY, MACOS, IOS, ANDROID, SPYING, PHOTO, STEGANOGRAPHY	RATs, Ransomware, Spyware

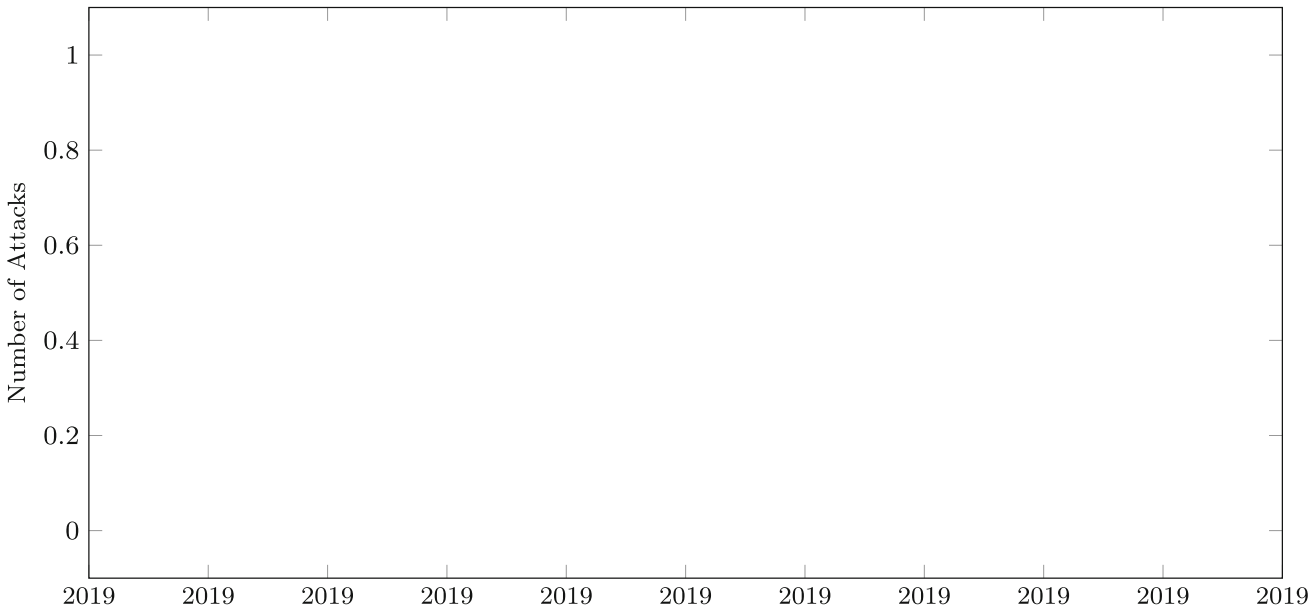

Fig. 4 Cyber security incidents 2019–2022

the minimum requirement. Nevertheless, this was impossible due to the lack of accurate reporting of cyber security events before 2019.

Given the dataset of incidents covering 2019 to 2022, described in Table 1b, we extracted the monthly statistics per topic and calculated the volume fluctuation per month, visualised in Fig. 4.

By providing the time series of cyber attack volumes to a set of forecasting models, we evaluated the most effective ones for trend prediction purposes, given the limitations mentioned above. We used 90% of the dataset to train the models and 10% for testing. We aimed to use the models trained with statistics of 2019 till the first semester of 2022 to predict incidents of the second semester of 2022 and validate them with the actual incidents that took place around the same period. Rather than aiming towards exact prediction matching for each month, we focus our effort on predicting the incident volume trends per period as follows: *upward* (↑), *slightly upward* (↗), *stable (-)*, *slightly downward* (↘), and *downward* (↓). The trend is calculated as the difference between two consecutive periods' average volume of incidents (Algorithm 1). In our case, the period used was six months.

Figure 5 maps the predicted values per month for the second semester of 2022 against the real volume of incidents, while Figs. 6 and 7 present the various model evaluation metrics.

Then, using this dataset, we decided to forecast the trends for 2022 for all topics (Table 6) and then performed a dedicated forecast by filtering our dataset per sector. We focused

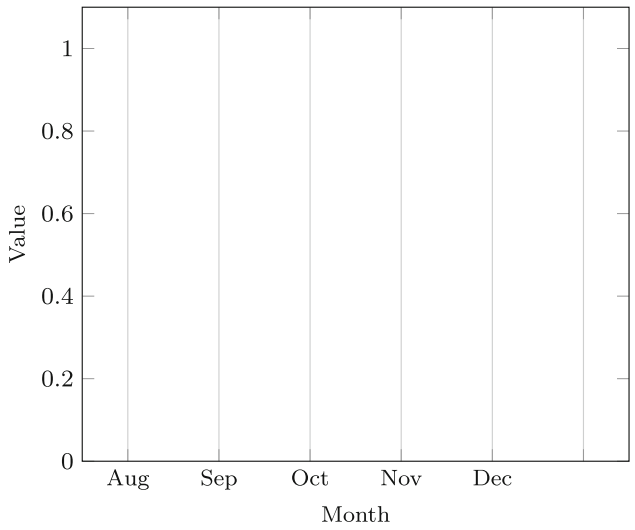

Fig. 5 Cyber security incidents prediction results (2nd Semester 2022)

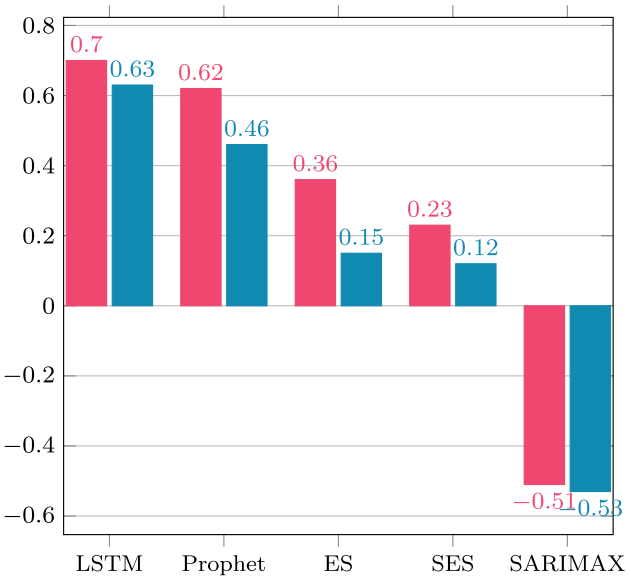

Fig. 6 R^2 Scores (Training, Testing Dataset) per Forecasting Model

Algorithm 1 Trend Value Labelling

```
Given TrendValue, ranging from negative to positive.
if TrendValue ≥ 6 then
    Label ← "Upward" {Trend is upward}
else if 3 < TrendValue < 6 then
    Label ← "Slightly upward" {Trend is slightly upward}
else if -3 ≥ TrendValue ≤ 3 then
    Label ← "Stable" {Trend is stable}
else if -6 < TrendValue < -3 then
    Label ← "Slightly downward" {Trend is slightly downward}
else
    Label ← "Downward" {Trend is downward}
end if
```

our efforts on four sectors, namely energy, healthcare, government, and finance, as seen in Table 7, using the forecasting algorithms side by side to ease the comparison of the results. This approach can help us generate tailored leads and forecasts and, as a result, more focused CSE scenarios per sector.

Four forecasting methods, namely LSTM, Prophet, ES, and SES, returned positive R^2 values, with LSTM being the fittest algorithm, closely followed by Prophet. The R^2 values of ES and SES, while positive, show that these models do not explain the variability of the response data around the mean; therefore, the predictions cannot be considered accurate. On the contrary, while not ideal, the range of R^2 values of both LSTM and Prophet implies a good fitness of the models. Finally, the negative values of SARIMAX clearly illustrate that this method is unsuitable for our dataset.

Based on the fact that LSTM had low MAE, MAPE, and MAD scores, it can be considered the most suitable algorithm for this task. This is validated by the fact that LSTM was one of the best in trend prediction when forecasting future volumes of the five topics per sector, as presented in Table 7. Similarly, Prophet had low MAE, MAPE, and MAD scores and satisfactory trend predictions when forecasting future volumes for the five topics per sector. Therefore, it can also be considered suitable for the task. SARIMAX, beyond the negative R^2 values, had high MAE, MAPE, and MAD values, validating that it cannot be used for accurate forecasting or trend prediction in our dataset.

From the other two methods with low R^2 values, ES had high MAE, MAPE and MAD scores, which was mapped in the poor performance when forecasting future incident volumes for the five topics per sector. Finally, even if SES did not fit well based on the R^2 values, it had the lowest MAE, MAPE and MAD scores and its performance was very satisfying for forecasting future incident volumes for the five topics per sector. We interpret the above as an indication that, while the predictions were not very accurate, the model was able to fit enough to identify the trends. For the rest of the steps in our proposed methodology, we will use the top three performing algorithms (Prophet, LSTM, SES) for further evaluation and comparison while dismissing ES and SARIMAX due to low performance in the selected metrics. When implementing our methodology in real life, the use of a single forecasting algorithm is adequate to provide the wanted predictions.

3.7 AI-driven prompt generation (text synthesis)

Our prompt generation process focuses on creating textual prompts based on a set of specific requirements by mixing user input with Current Topic Trends (CTTs) and Future Topic Trend Predictions (FTTPs) extracted from previous steps. The user provides a dataset of incidents and the sector of interest (optional) as input to our methodology. The user can also define the forecasting model to be used. Our AI-

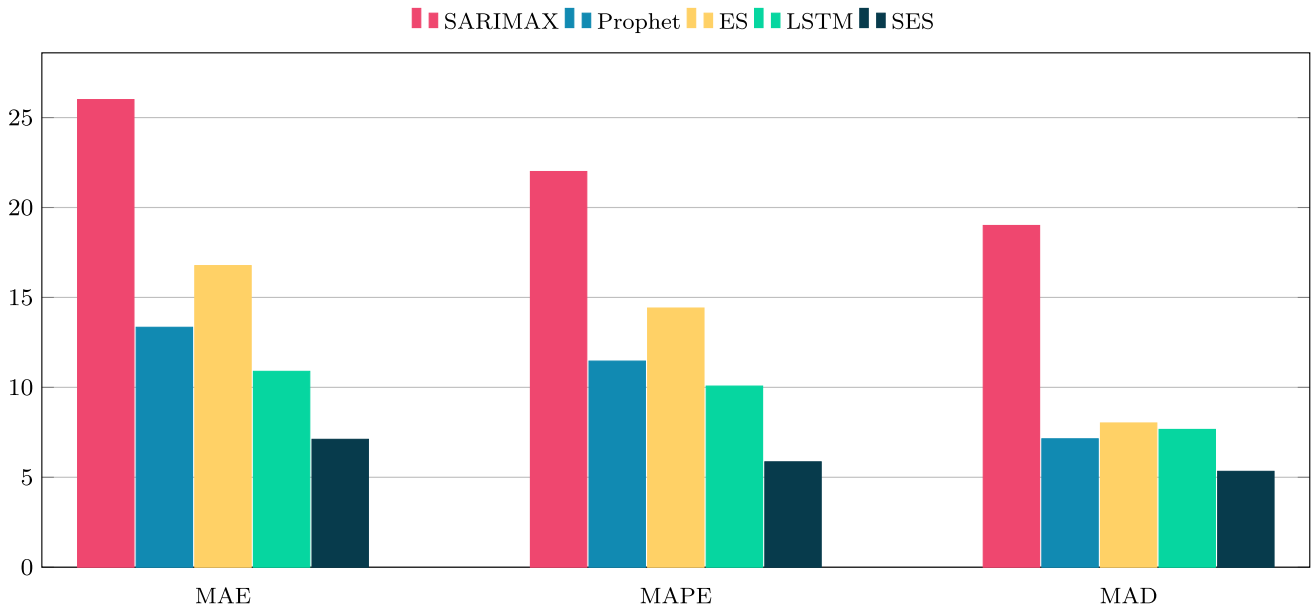

Fig. 7 MAE, MAPE, and MAD score per forecasting model

Table 6 Trend forecasting results for the 2nd Semester of 2022

	Actual Trend 2022	SARIMAX	Prophet	ES	SES	LSTM
All data-topic agnostic	–	↓	↓	↓	↓	–
RATS, ransomware, trojans	↘	↓	–	–	↘	–
Network_System_Exploit	↓	–	↓	↓	↓	↓
Phishing, Social_Eng	↘	↑	↑	↑	↘	↘
Personal_Data_Breach	↘	↘	↓	↑	↘	↘
Card_Fraud_Online_Theft	↘	–	↓	↑	–	–

driven methodology extracts the following information from the dataset based on the user input:

- **Current Topic Trends (CTTs)**: The top ten keywords in volume for the following tag categories: [ATTACK TYPE], [ATTACKER NAME], [VULNERABILITY], [ASSETS], [TECHNOLOGY], [MALWARE NAME], [MALWARE TYPE] for a given [SECTOR] over the course of a year. This is achieved by executing steps 1 and 2 of our methodology.
- **Future Topic Trend Predictions (FTTPs)**: A set of keywords that stem from previously defined topics (Table 5) and that exhibit a clear upward trend based on our trend prediction methodology by using the user selected forecasting algorithm. This is achieved by executing steps 1, 2, 3 and 4 of our methodology.

A textual prompt template consisting of four sections (with fixed structure along with dynamic sections) is then used to generate prompts that will form the basis of a CSE scenario consisting of the fifth and final step of our methodology.

An example of a prompt template can be seen in Fig. 8

The resulting output can then be edited or used as input to AiCEF, and a CSE incident will be generated, ready to be incorporated in an exercise scenario, as visualised in Fig. 9. Our methodology can produce meaningful, sectorial-focused prompts and scenarios with trend prediction incorporated seamlessly into the prompt without the need for cyber security expertise from the user.

4 Evaluation methodology and results

4.1 Evaluation methodology

To evaluate and measure the effectiveness of our proposed framework and the underlying methodology, we developed a case study focusing on the energy sector. The energy sector, present in both NIS and NIS2 [14] Directives, was selected due to its high critical impact on other industries if affected, but also due to its unique threat landscape that attracts a specific set of threat actors with unique motivation, Tactics, Techniques, and Procedures (TTPs). To this end, we followed a three-step methodology.

Table 7 Sectorial trend forecasting results for 2022

Sector	Actual Trend 2022	LSTM	Prophet	SARIMAX	ES	SES
Energy						
Sectorial Incidents	–	–	–	↗	↗	–
Personal_Data_Breach	–	–	–	↗	↗	–
Phishing_Social_Eng	–	–	–	–	–	–
Netwok_System_Exploit	–	–	↗	↗	↗	–
Card_Fraud_Online_Theft	–	–	↗	↗	↗	–
Rats_Ransomware_Trojans	–	–	–	–	↗	–
Healthcare						
Sectorial Incidents	–	–	–	↗	–	–
Personal_Data_Breach	–	–	–	↗	–	–
Phishing_Social_Eng	–	–	–	–	–	–
Network_System_Exploit	–	–	–	↗	↗	–
Card_Fraud_Online_Theft	–	–	–	↗	↗	–
Rats_Ransomware_Trojans	–	–	–	–	–	–
Government						
Sectorial Incidents	↓	–	–	–	↓	–
Personal_Data_Breach	↓	–	–	–	↓	–
Phishing_Social_Eng	↓	–	–	–	↓	↓
Network_System_Exploit	–	–	–	↑	↗	–
Card_Fraud_Online_Theft	↓	–	–	–	↓	↓
Rats_Ransomware_Trojans	↘	–	–	↗	↓	↓
Financial						
Sectorial Incidents	↓	–	↘	↓	↓	↓
Personal_Data_Breach	↓	–	↘	↓	↓	↓
Phishing_Social_Eng	↓	↓	↘	↓	↓	↓
Network_System_Exploit	–	–	–	–	–	–
Card_Fraud_Online_Theft	↓	–	↘	↓	↓	↓
Rats_Ransomware_Trojans	↓	–	↓	↓	↓	↓

1. **Prompt generation** We assigned exercise planners (EPs) with varying expertise to generate a set of textual prompts as input to our AiDPG methodology. Two groups of experts were used: experts in the field of cyber exercise development and experts with knowledge of the specific sector but with little or no experience in the exercise development role. The choice of creating a group of human experts consisting of the two subgroups mentioned above was made so that the two could combine their expertise and generate an exercise with both sectorial focus and technical value for the participant. Our approach of creating mixed groups of exercise planners with sectorial expertise and exercise planning background is not new; in fact, it is promoted by national cyber security agencies such as ANSII.[5] Consequently, three sets of prompts were generated by humans and a fourth one using our AiDPG methodology:
   - Prompt Set 1: The Exercise experts prompt set, created by the experts in CSE development.
   - Prompt Set 2: The Sector experts prompt set, created by experts in the sector examined (ENERGY).
   - Prompt Set 3: The Human experts prompt set, which is the combination of sets 1 & 2.
   - Prompt Set 4: The AiDPG prompt set was generated purely using our methodology focusing on the sector of interest (ENERGY) with forecasting enabled.

   AiDPG (no forecasting) prompt set was also generated as a subset of Prompt Set 4 and will be used only for limited comparisons against other prompt sets.
2. **Prompt Evaluation** All four sets of prompts were lexicologically analysed to identify keyword patterns and biases, using AiCEF for annotation, leading to the following tag sets:

[5] https://www.ssi.gouv.fr/uploads/2021/09/anssi-guide-organising_a_cyber_crisis_management_exercise-v1.0.pdf

Prompt Template:

- **Section 1: Attack Type and Sector**
  - Example: ATTACK TYPE in the SECTOR.
- **Section 2: Trending TTPs and Impact based on STI**
  - Example: Attackers ATTACKER NAME used VULNERABILITY to affect ASSETS.
  - The malware employed is MALWARE NAME MALWARE TYPE.
  - Technologies used or abused: TECHNOLOGY.
- **Section 3: Future Trending Topics** (*This section can be omitted if not enough data are available to perform forecasting)
  - Use NUM of the following trending elements: SECTOR TOPIC[0] ITEMS
- **Section 4: Styling and Formatting**
  - Start the prompt with: ["*Generate a cyber security incident describing an* "]. Prompt size: [200] words max

Example Prompt:

Generate a cyber security incident describing a BREACH in the ENERGY SECTOR. Attackers CONTI used CVE-2022-24002 to affect PASSWORDS, DATA. The malware employed is CONTI RANSOMWARE. Technologies used or abused: WINDOWS. The following trending elements were used in the incident: PHISHING, SPEAR-PHISHING, MALWARE.

Fig. 8 The form of our template above and an example prompt, below

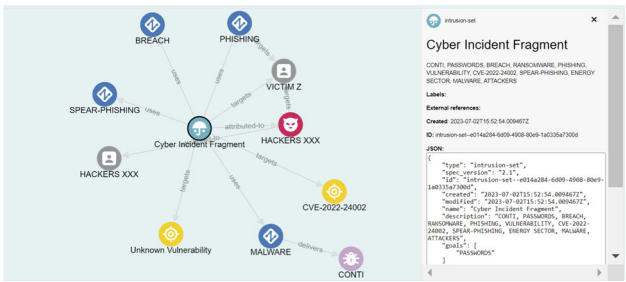

Fig. 9 STIX2.1 output generated by AiCEF

- "Exercise experts" Tag set: Prompt Set 1 (The Exercise experts prompt set) was annotated, forming this tag set.
- "Sector experts" Tag set: Prompt Set 2 (The Exercise experts prompt set) was annotated, forming this tag set.
- "Human Experts" Tag set: Prompt set 3 (All Human experts prompt set) was annotated, forming this tag set.

Prompt Specs: Generate a CSE prompt for a cyber awareness exercise. The prompt should not be longer than 600 words and should include as many technical details as possible related to a cyber security incident. The company which will use the scenario is an **Energy Service Provider** and all its employees can be potential Players. The incident generated should focus on the existing threat landscape and should train employees to prepare for current or future threat trends relevant to the designated sector in an exercise that will take place in 2023.

Fig. 10 Task definition

- "AiDPG Energy" Tag set: The Prompt set 4 (AiDPG prompt set) was annotated, forming this tag set.
- "Energy Threat Landscapes" Tag set: A collection of Sectorial Threat Landscapes (STLs) were annotated, forming this baseline Tag set.

Using these five tag sets, we performed a number of comparisons in an attempt to identify the prompt maturity and coverage of sector-oriented threat landscapes by comparing prompts generated using our methodology to those generated by human experts.

3. **Trend Forecasting Evaluation**
   Finally, to further evaluate and verify our trend prediction and forecasting findings, we surveyed the two groups of experts to test our forecasts against human expertise and intuition.

In the following paragraphs, we drill down to each step and present our evaluation results.

4.2 Prompt generation

Twenty-six cyber security experts were selected to generate five (5) individual CSE scenarios prompts each, according to the high-level requirements and specifications, as illustrated in Fig. 10.

All experts have cyber security backgrounds, and their skill sets resemble those of a Chief Information Security Officer (CISO) or similar. For the cyber exercise expert group, we invited several planners with high expertise in cyber security exercise management, mixed ethnicity, and focus sectors. More precisely, we invited members of a Cyber Awareness Expert Group, [6] from which 12 exercise planners responded. The sectorial experts group had a cybersecurity background for the energy sector (Members of the European Network of Transmission System Operators for Gas or European Network of Transmission System Operators for Electricity) and acted as CISOs for their organisations. In total, 14 sectorial experts responded. The demographic distribution of all experts is illustrated in Table 8.

[6] https://www.enisa.europa.eu/topics/cybersecurity-education/ad-hoc-working-group-awareness-raising

Table 8 Demographics of the experts

(a) Countries of origin of the experts

Country	#	Country	#
France	3	Luxembourg	1
Netherlands	3	Lithuania	1
Czechia	2	Austria	1
Spain	2	Italy	1
Belgium	2	Finland	1
Greece	2	Ireland	1
Denmark	2	Portugal	1
Croatia	1	Poland	1
Germany	1	Total	**26**

(b) Sector that experts are working in

Sector	#
Energy	14
Public Administration	5
ICT	2
Transport	2
Law Enforcement	2
Education	1

Not all of the experts provided the maximum prompt number of five (5) as requested, so we collected only fifty-five (55) prompts per group, 110 human prompts in total. All experts provided at least three prompts.

The prompt specs were inputted in our methodology, and an AiDPG prompt set was generated, focusing on the energy sector concluding the prompt generation task.

4.3 Prompt evaluation

To evaluate the sets of generated prompts, we applied keyword extraction using AiCEF.

Two types of evaluation metrics were used:

- Prompt maturity, using AiCEF's maturity calculation algorithm, to evaluate the richness of the prompt in terms of exercise-related objects.
- Tag coverage of the human and AI-assisted tag sets against that of a sectorial baseline. This metric helps evaluate the relevance of the prompts generated.

4.3.1 Prompt maturity

Based on AiCEF's maturity calculation algorithm, the maximum maturity score for any text can be 150 points, indicating that the text analysed has the maximum richness of information, covering all MLCESO tags in Table 3.

Table 9 Maturity scores

Label Name	Prompts	Maturity Score
Sector Experts	55	50,41
Exercise Experts	55	73,94
AiDPG (no forecasting)	55	135
AiDPG	55	150

Table 9 illustrates the average maturity score per group of prompts. The average maturity score of AiDPG (150) (even with no forecasting (135)) clearly exceeds the scores of any of the two human expert groups.

Indeed, these scores provide sufficient proof that our sector-focused AiDPG methodology can help human CSE planners improve their prompts and effectively exercise scenarios. More precisely, a 50.4% improvement for the exercise experts and an improvement of 66.67% for sector experts can be achieved in terms of prompt maturity.

4.3.2 Tag coverage

After analysing and comparing the accuracy of the lexicological elements of the prompt sets generated by exercise experts, sector experts (energy experts) and their combination along with our AiDPG methodology, we defined a baseline set of tags. The baseline tag set was generated by extracting tags (see Table 3) from acknowledged STL reports. More precisely, we used three STLs to evaluate our methodology: the Deloitte Oil and Gas (Energy) Threat Landscape, [7] the Energy Sector 2022 Threat Landscape by CITALID and Sekoia.io [8] and the Electric Avenue report by BlueViv & Neurosoft.[9] These threat landscapes cover the broader energy sector, using cyber security incidents from as early as 2017 up to early 2023. The unique tags per report are visualised in Fig. 11.

The analysis of the following tag sets can be visualised in Fig. 12.

- "Exercise experts" Tag set
- "Sector experts" Tag set
- "Human Experts" Tag set
- "AiDPG Energy" Tag set
- "Energy Threat Landscapes" Tag set

[7] https://www2.deloitte.com/us/en/pages/risk/articles/oil-and-gas-sectors-threat-landscape.html

[8] https://blog.sekoia.io/the-energy-sector-2022-cyber-threat-landscape/

[9] https://neurosoft.gr/the-electric-avenue-an-overview-of-the-energy-sectors-threat-landscape/

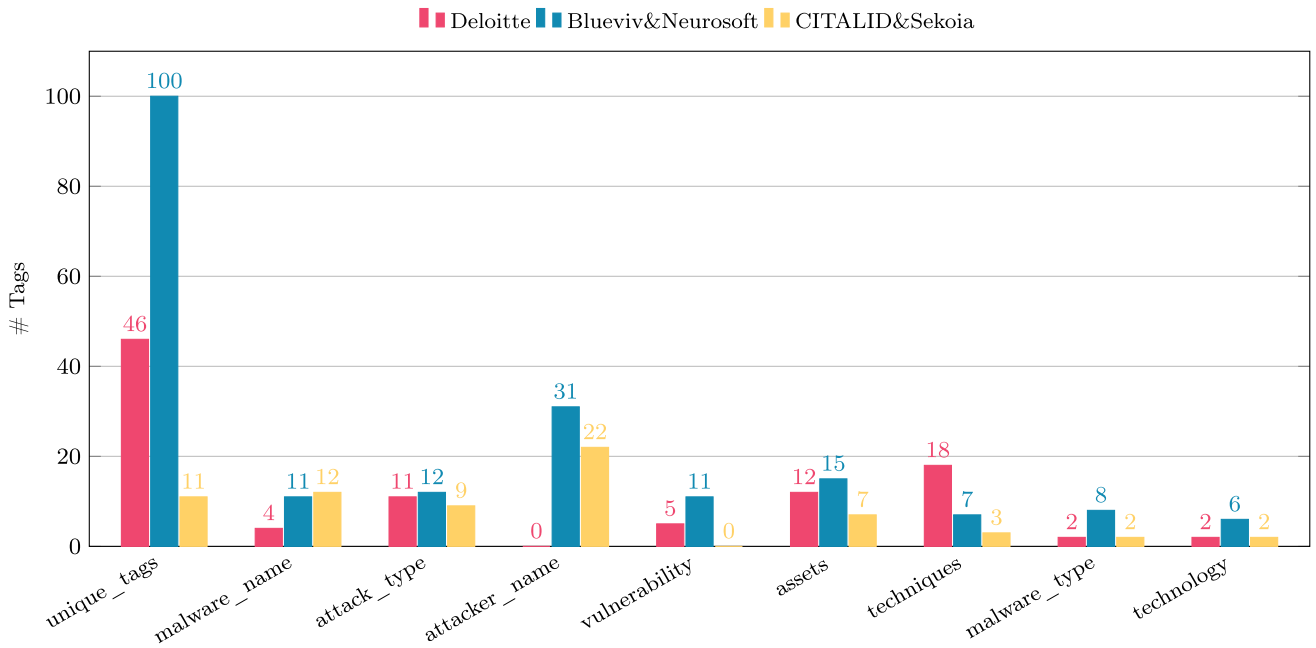

Fig. 11 Unique annotation tags values per energy landscape report

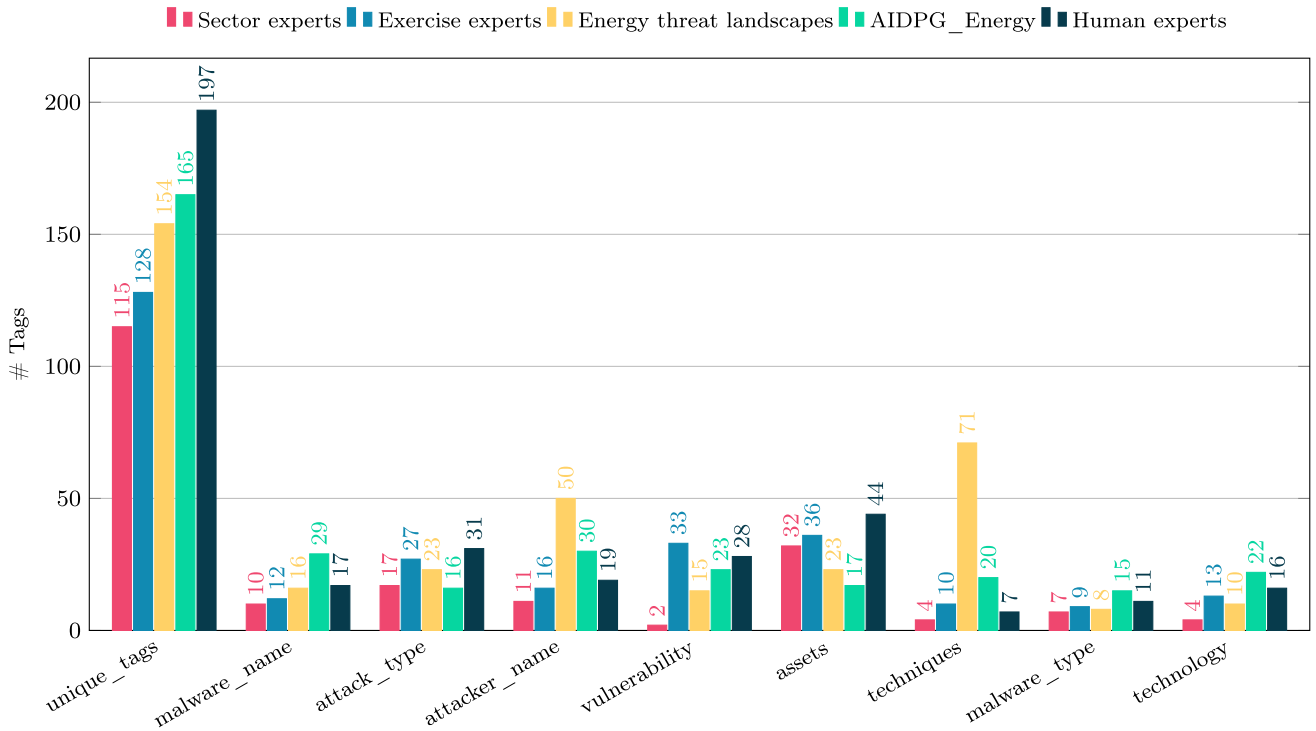

Fig. 12 Unique annotation tag volume comparison

Finally, the overlapping unique tags for "ENERGY THREAT LANDSCAPES", "HUMAN EXPERTS", and "AiDPG ENERGY" are represented in Fig. 13 providing a better visual representation of the coverage between Prompt and eventually Tag sets.

To verify the superiority of AiDPG methodology over human-expert-generated prompts, we use Jaccard similarity to calculate the similarity between two sets of unique phrases based on the number of common phrases and the total number of distinct phrases in the sets.

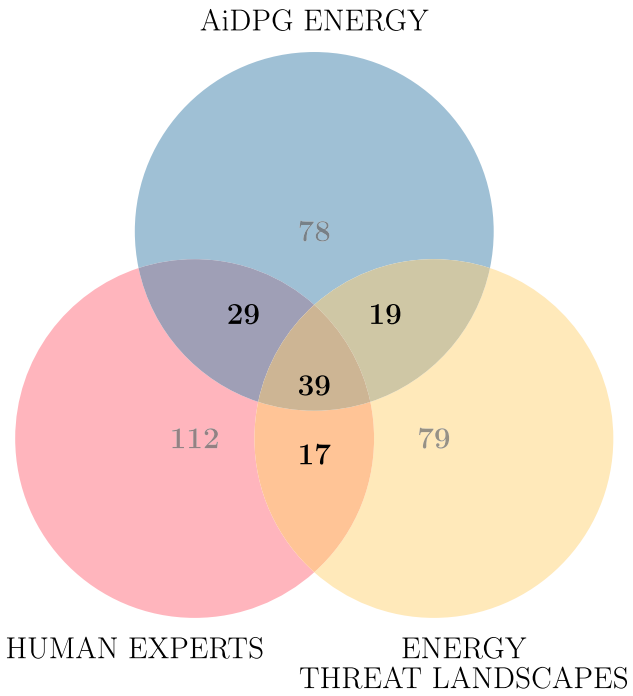

Fig. 13 Unique tags of STLs versus Human Experts and AiDPG

The Jaccard similarity coefficient (JSC) is defined as the size of the intersection of the sets divided by the size of the union of the sets.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where:

J(A, B) represents the Jaccard similarity coefficient (JSC) between sets A and B.

$|A \cap B|$ represents the size (cardinality) of the intersection of sets A and B.

$|A \cup B|$ represents the size (cardinality) of the union of sets A and B.

After comparing the Jaccard similarity, it is clear that the prompts generated with our AiDPG methodology are more relevant to the set generated when combining all energy threat landscapes:

- JSC for "AiDPG Energy" and "Energy Threat Landscapes" with value: 0.22
- JSC for "Human Experts" and "Energy Threat Landscapes" with value: 0.19
- JSC for "Sector Experts" and "Energy Threat Landscapes" with value: 0.16
- JSC for "Exercise Experts" and "Energy Threat Landscapes" with value: 0.17

Focusing on which group of experts can benefit more from our AiDPG methodology, we calculate the unique annotation tag categories intersection per group and measure the number of tags that are present in both "Energy Threat Landscapes" and "AiDPG Energy" but not in the individual human generated sets. By doing this, we assume that, at a minimum, the content of the "Energy Threat Landscapes" that is also part of the "AiDPG Energy" set can improve the prompt and be relevant.

Let A represent the set "Energy Threat Landscapes", B represent the set "AiDPG Energy", and C represent a human-generated set. The number of tags in the intersection of A and B but not in C can be calculated as follows:

$$\text{Benefit} = \|A \cap B\| - \|A \cap B \cap C\|$$

$$\text{BenefitPercentage} = \frac{\text{IntersectionCount}}{\|A\|} \times 100\%$$

Using the formula above for the annotation tags in Table 3, we can calculate the individual benefit percentages that the two human groups of experts could achieve using our methodology, Fig. 14.

The results underline the effectiveness of our methodology since both expert groups can benefit from at least 18.83% more relevant tags proposed by AiDPG, not selected by them but present in the "Energy Threat Landscapes" set.

4.4 Trend forecasting evaluation

To compare our sectorial trend forecasting results, a survey was run against our pool of specialised planners. The survey aimed to collect insights on the prioritisation of cyber threats and assets/resources based on both established threat landscape data and personal research, with a focus on preparing for incidents that are anticipated to occur in the coming months in the Energy sector. The ENISA Threat Landscape 2022[10] was used to extract the following eight prime cyber threats and assets affected and link them directly to the five types of incidents.

Cyber Threats: Ransomware, Malware, Disinformation/ misinformation, Threats against availability—(DOS, DDOS), Phishing & Social engineering threats, Threats against data (data breaches & leaks), Supply-chain attacks, Physical threats with security impact.

Assets Affected: Company Data, Email or other Credentials, Personal Information, Website, Network or other Infrastructure, Sensitive Personal Information, Monetary Funds, Reputation.

The planners were asked to provide feedback on the following questions:

[10] https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

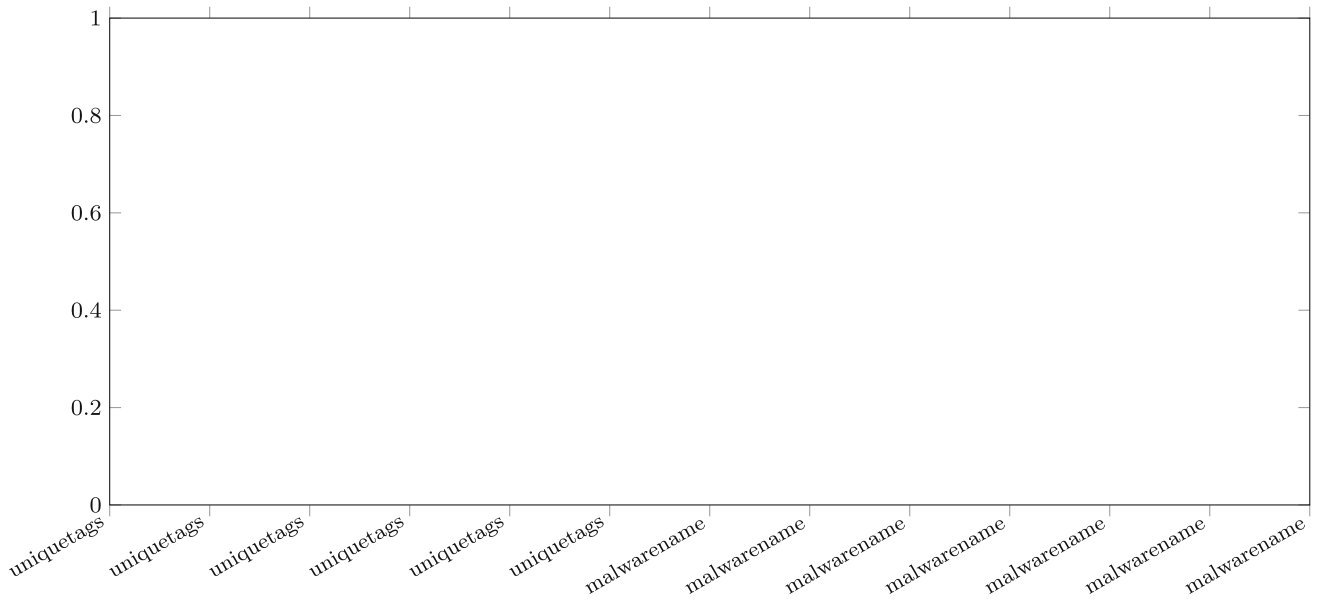

Fig. 14 Benefit percentage per tag and expert group

1. What is your trend prediction on the overall volume of cyber incidents in 2023 compared to 2022 affecting the energy sector?
2. What is your trend prediction for the following types of incidents in 2023 compared to 2022 affecting the energy sector?
   - (Personal) Data Breaches
   - Phishing & Social Engineering
   - Network & System Exploitation
   - Payment Card Fraud & Online Theft
   - Remote Access Trojans (RATs), Ransomware, Spyware

The possible answers were: *upward* (↑), *slightly upward* (↗), *stable (-)*, *slightly downward* (↘), and *downward* (↓)

We then implemented our proposed trend forecasting methodology for energy-related incidents for 2023 using the top three scoring models (Prophet, SES, and LSTM). A comparison of our results against the expectations of the two expert groups can be found in Tables 10 and 11, where both the normalised forecasted score (as per Algorithm 1) and trend is presented.

Regarding forecasting the overall sectorial incident trends, the two groups of humans agreed that there would be a "Slightly Upward" trend in 2023, a result also captured by our LSTM model. When comparing human foresight on sectorial incident trends per topic, we notice that the two expert groups do not align. Focusing more on the foresight of sectorial experts, we see that both Prophet and LSTM are making trend predictions that are close to those of humans. The results are satisfying in that, in most cases, our methodology did not produce the opposite results to the humans' foresight. Yet, our methodology is based on incident volumes and cannot replicate human foresight consisting of broader knowledge aspects that can contribute to a change trends like geopolitics. Only the actual data of cyber incidents against the energy sector in 2023 can prove whether our trend forecast or human insight was more accurate in the predictions made, a task to be performed in future work.

Table 10 Sectorial trend forecasting results for 2023

Forecast Type	Overall Energy Attacks trend
Prophet	7, Upward Trend (↑)
SES	0, Stable (–)
LSTM	3.17, Slightly Upward (↗)
SECTOR EXPERTS	3.9, Slightly Upward (↗)
EXERCISE EXPERTS	3.75, Slightly Upward (↗)

5 Conclusions and future work

CSEs can be a powerful tool in improving the organisation's preparedness against the ever-changing threat landscape. However, experts' shortages, timeliness, and relevance of the developed CSEs require novel solutions. To address this challenge, our proposal revolves around incorporating AI-driven sectorial threat landscaping and forecasting into existing exercise generation frameworks like AiCEF. This integration aims to identify emerging and significant threats and predict their potential impact in various sectors. By infusing these forecasts into creating cyber security exercise prompts using AiDPG and, eventually, scenarios, organisations can

Table 11 Sectorial forecast results per topic for the year 2023

Forecast Type	Card Fraud & Online Theft	Network System Exploit	Personal Data Breach	Phishing & Social Eng	Rats, Ransomware, Trojans
Prophet	4, ↗	9, ↑	7, ↑	5, ↑	4, ↗
SES	1, –	3, –	0, –	0, –	0, –
LSTM	2.1, –	3.5, ↗	3.1, ↗	1.9, –	2.4, –
SECTOR EXPERTS	2.71, –	4.29, ↑	3, –	4.64, ↑	4.15, ↑
EXERCISE EXPERTS	2.08, ↘	3.58, ↗	2.41, ↘	4.75, ↑	4.58, ↑

achieve more precise simulations of real-world cyber security incidents and better evaluate their preparedness against emerging threats.

Our proposed methodology identified the gap in context relevant CSE scenarios generated by sectorial cyber security experts and seasoned exercise planners. By introducing a novel methodology, we can provide tangible quantitative added value in CSE scenario development by tackling the problems of relevance and timeliness that are independent of the EP's experience. More precisely, through our AI-driven topic extraction model and by clustering sector-specific annotation tags, we can compare various forecasting models and accurately identify emerging (sectorial) cyber threats for a given set of cyber security incidents. To this end, we compared SARIMAX, ES, Simple Exponential Smoothing (SES), Prophet, and Long Short-Term Memory (LSTM) modelling using Recurrent Neural Networks (RNN). Our experiments illustrated that for this dataset, LSTM and Prophet were the two fittest algorithms for this task.

Regarding prompt maturity, the sectorial-focused AiDPG methodology can indeed assist human CSE planners in improving their prompts and effectively the exercise scenarios. More specifically, a 50.4% improvement for the exercise experts and a 66.67% improvement for the sector experts can be achieved in terms of prompt maturity, generating not only richer but also more sector-focused scenarios.

After calculating the Jaccard similarity, it is clear that the prompts generated by our AIDPG methodology are more relevant to the set generated when combining all energy threat landscapes than the set generated by all human experts combined:

- "AiDPG Energy" and "Energy Threat Landscapes" with value: 0.22
- "HUMAN Experts" and "Energy Threat Landscapes" with value: 0.19.

Even more, by leveraging our user surveys to measure the perceived relevance of CSE with AI-driven threat intelligence compared to traditional exercise generation, we managed to quantify the value of our methodology. The results underline the effectiveness of our methodology since both expert groups surveyed, Exercise Experts and Sector Experts, can improve by at least 18.83% the prompt and effective scenario relevance to a specific sector.

Limitations to our methodology have been identified regarding the volume and time distribution of past incidents needed in order to form accurate forecasting. For Prophet specifically, 3 years of observations is the minimum data to predict a fourth year in sequence. We consider the amount of observation of three years to be the minimum for satisfactory predictions to incorporate forecasting results in our proposed methodology. Further to the above, forecasting algorithms cannot replace human foresight based on the broader knowledge of a domain that leads human experts to expect trend changes. We believe that the use of Large Language Models will contribute to the improvement of our methodology towards bridging this gap.

In the future evolution of our work, we will aim to fully automate threat landscape generation procedures by processing any given data set. In the short term, we plan to collect cyber incident statistics for 2023. We will then properly evaluate the results of our presented case study on the energy sector to identify if our trend forecast methodology was more accurate in predicting trends against the human experts' insight. While our methodology proves to be relatively efficient in focused prompt and scenario generation, more tests on non-expert humans should be conducted to identify further benefits in the CSE creation and customisation process. To verify our hypothesis, we plan to embed the AiDPG methodology in AiCEF to help automatically produce CSE scenarios and measure the adoption and satisfaction of the final output by experts in the domain.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10207-024-00860-w.

Acknowledgements This work was supported by the European Commission under the Horizon Europe Programme, as part of the projects CyberSecPro (https://www.cybersecpro-project.eu) (Grant Agreement no.101083594) and LAZARUS (https://lazarus-he.eu/) (Grant Agreement no. 101070303). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

Funding Open access funding provided by HEAL-Link Greece.

Data Availability The data used in this work is available at https://github.com/alexzacharis/AI-driven-Threat-Intelligence-and-Forecasting.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Acarturk, C., Sirlanci, M., Balikcioglu, P.G., Demirci, D., Sahin, N., Kucuk, O.A.: Malicious code detection: run trace output analysis by lstm. IEEE Access **9**, 9625–9635 (2021)
2. Almahmoud, Z., Yoo, P.D., Alhussein, O., Farhat, I., Damiani, E.: A holistic and proactive approach to forecasting cyber threats. Sci. Rep. **13**(1), 8049 (2023)
3. ANSII: Organising a cyber crisis management exercise (2021). https://www.ssi.gouv.fr/guide/organising-a-cyber-crisis-management-exercise/
4. Armstrong, J.S., Collopy, F.: Error measures for generalizing about forecasting methods: empirical comparisons. Int. J. Forecast. **8**(1), 69–80 (1992)
5. Augustine, T., Dodge, R.C., et al.: Cyber defense exercise: meeting learning objectives thru competition. In: Proceedings of the 10th Colloquium for Information Systems Security Education (2006)
6. Bakdash, J.Z., Hutchinson, S., Zaroukian, E.G., Marusich, L.R., Thirumuruganathan, S., Sample, C., Hoffman, B., Das, G.: Malware in the future? Forecasting of analyst detection of cyber events. J. Cybersecur. **4**(1), tyy007 (2018)
7. Bilge, L., Han, Y., Dell'Amico, M.: Riskteller: predicting the risk of cyber incidents. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1299–1311 (2017)
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
9. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. John Wiley & Sons (2015)
10. Carriegos, M.V., Fernández-Díaz, R.Á.: Towards forecasting time-series of cyber-security data aggregates. In: Herrero, Á., Cambra, C., Urda, D., Sedano, J., Quintián, H., Corchado, E. (eds.) 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020), pp. 273–281. Springer International Publishing, Cham (2021)
11. Couce-Vieira, A., Insua, D.R., Kosgodagan, A.: Assessing and forecasting cybersecurity impacts. Decis. Anal. **17**(4), 356–374 (2020)
12. ENISA: National Exercise—Good Practice Guide (2009). https://www.enisa.europa.eu/publications/national-exercise-good-practice-guide
13. ENISA: ENISA CYBERSECURITY THREAT LANDSCAPE METHODOLOGY (2022). https://www.enisa.europa.eu/publications/enisa-threat-landscape-methodology/
14. European Commission: Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) (Text with EEA relevance) (2022). https://eur-lex.europa.eu/eli/dir/2022/2555
15. Fan, S., Wu, S., Wang, Z., Li, Z., Yang, J., Liu, H., Liu, X.: Aleap: attention-based lstm with event embedding for attack projection. In: 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC), pp. 1–8. IEEE (2019)
16. Fang, Z., Xu, M., Xu, S., Hu, T.: A framework for predicting data breach risk: leveraging dependence to cope with sparsity. IEEE Trans. Inf. Forensics Secur. **16**, 2186–2201 (2021)
17. Furtună, A., Patriciu, V.V., Bica, I.: A structured approach for implementing cyber security exercises. In: 2010 8th International Conference on Communications. IEEE, pp. 415–418 (2010)
18. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint (2023) arXiv:2303.15056
19. Gogineni, K., Derasari, P., Venkataramani, G.: Foreseer: efficiently forecasting malware event series with long short-term memory. In: 2022 IEEE International Symposium on Secure and Private Execution Environment Design (SEED). IEEE, pp. 97–108 (2022)
20. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
21. Goyal, P., Hossain, K., Deb, A., Tavabi, N., Bartley, N., Abeliuk, A., Ferrara, E., Lerman, K.: Discovering signals from web sources to predict cyber attacks (2018). arXiv preprint arXiv:1806.03342
22. Granåsen, M., Andersson, D.: Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. Cogn. Technol. Work **18**(1), 121–143 (2016)
23. Green, A., Zafar, H.: Addressing emerging information security personnel needs. A look at competitions in academia: Do cyber defense competitions work? AMCIS 2013 Proceedings 1, 257 (2013)
24. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
25. Husák, M., Kašpar, J.: Aida framework: real-time correlation and prediction of intrusion detection alerts. In: Proceedings of the 14th International Conference on Availability, Reliability and Security, pp. 1–8 (2019)
26. Husák, M., Komárková, J., Bou-Harb, E., Čeleda, P.: Survey of attack projection, prediction, and forecasting in cyber security. IEEE Commun. Surv. Tutor. **21**(1), 640–660 (2018)
27. Husák, M., Bartoš, V., Sokol, P., Gajdoš, A.: Predictive methods in cyber defense: current experience and research challenges. Future Gener. Comput. Syst. **115**, 517–530 (2021)
28. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts (2018)
29. Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S.: A state space framework for automatic forecasting using exponential smoothing methods. Int. J. Forecast. **18**(3), 439–454 (2002)
30. Hyndman, R.J., Koehler, A.B., Ord, J.K.: Forecasting with Exponential Smoothing: The State Space Approach. Springer Science & Business Media (2008)
31. ISO Central Secretary: Societal security—guidelines for exercises. Standard ISO22398:2013, International Organization for Standardization, Geneva, CH (2013). https://www.iso.org/standard/50294.html

32. Jansen, B.J., Sg, Jung, Salminen, J.: Employing large language models in survey research. Nat. Lang. Process. J. **4**(100), 020 (2023)
33. Kick, J.: Cyber Exercise Playbook. Tech. rep, MITRE CORP BEDFORD MA (2014)
34. Kim, J., Kim, J., Thu, H.L.T, Kim, H.: Long short term memory recurrent neural network classifier for intrusion detection. In: 2016 International Conference on Platform Technology and Service (PlatCon). IEEE, pp. 1–5 (2016)
35. Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W.: Applied Linear Statistical Models. McGraw-hill (2005)
36. Liu, X., Liu, J.: Malicious traffic detection combined deep neural network with hierarchical attention mechanism. Sci. Rep. **11**(1), 12,363 (2021)
37. Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., Liu, M.: Cloudy with a chance of breach: Forecasting cyber security incidents. In: 24th USENIX Security Symposium (USENIX Security 15), pp. 1009–1024 (2015)
38. MacIntyre, R.: Penn treebank tokenizer (sed script source code) (1995)
39. Malik, J., Akhunzada, A., Bibi, I., Imran, M., Musaddiq, A., Kim, S.W.: Hybrid deep learning: an efficient reconnaissance and surveillance detection mechanism in sdn. IEEE Access **8**, 134,695-134,706 (2020)
40. Md Azam, M.N., Ramli, N.A.: Reported malicious codes incident within malaysia's landscape: time series modelling and a timeline analysis. Int. J. Adv. Data Sci. Intell. Anal. **2**(2) (2022). https://amcs-press.com/index.php/ijadsia/article/view/65
41. Munkhdorj, B., Yuji, S.: Cyber attack prediction using social data analysis. J. High Speed Netw. **23**(2), 109–135 (2017)
42. Okutan, A., Werner, G., Yang, S.J., McConky, K.: Forecasting cyberattacks with incomplete, imbalanced, and insignificant data. Cybersecurity **1**, 1–16 (2018)
43. Okutan, A., Yang, S.J., McConky, K., Werner, G.: Capture: cyberattack forecasting using non-stationary features with time lags. In: 2019 IEEE Conference on Communications and Network Security (CNS). IEEE, pp. 205–213 (2019)
44. Patsakis, C., Lykousas, N.: Man vs the machine in the struggle for effective text anonymisation in the age of large language models. Sci. Rep. (2023)
45. Qin, X., Lee, W.: Attack plan recognition and prediction using causal networks. In: 20th Annual Computer Security Applications Conference. IEEE, pp. 370–379 (2004)
46. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 248–256 (2009)
47. Ren, K., Zeng, Y., Cao, Z., Zhang, Y.: Id-rdrl: a deep reinforcement learning-based feature selection intrusion detection model. Sci. Rep. **12**(1), 15,370 (2022)
48. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)
49. Salih, A., Zeebaree, S.T., Ameen, S., Alkhyyat, A., Shukur, H.M.: A survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection. In: 2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC). IEEE, pp. 61–66 (2021)
50. Sarker, I.H., Kayes, A., Badsha, S., Alqahtani, H., Watters, P., Ng, A.: Cybersecurity data science: an overview from machine learning perspective. J. Big Data **7**, 1–29 (2020)
51. Schepens, W.J., James, J.R.: Architecture of a cyber defense competition. In: SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483). IEEE, vol. 5, pp. 4300–4305 (2003)
52. Shen, Y., Mariconti, E., Vervier, P.A., Stringhini, G.: Tiresias: predicting security events through deep learning. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, pp. 592–605 (2018)
53. Sievert, C., Shirley, K.: Ldavis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 63–70 (2014)
54. Suciu, O., Nelson, C., Lyu, Z., Bao, T., Dumitraş, T.: Expected exploitability: predicting the development of functional vulnerability exploits. In: 31st USENIX Security Symposium (USENIX Security 22), pp. 377–394 (2022)
55. Sun, N., Zhang, J., Rimba, P., Gao, S., Zhang, L.Y., Xiang, Y.: Data-driven cybersecurity incident prediction: a survey. IEEE Commun. Surv. Tutor. **21**(2), 1744–1772 (2018)
56. Sun, N., Ding,. M., Jiang, J., Xu, W., Mo, X., Tai, Y., Zhang, J.: Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives. IEEE Commun. Surv. Tutor. (2023)
57. Tavabi, N., Goyal, P., Almukaynizi, M., Shakarian, P., Lerman, K.: Darkembed: exploit prediction with neural language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
58. Taylor, S.J., Letham, B.: Prophet: forecasting at scale. J. Open Source Softw. **3**(22), 651 (2017)
59. Törnberg, P.: Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning (2023). arXiv preprint arXiv:2304.06588
60. Velleman, P.F., Hoaglin, D.C.: Applications, Basics, and Computing of Exploratory Data Analysis. Duxbury Press (1981)
61. Werner, G., Yang, S., McConky, K.: Time series forecasting of cyber attack intensity. In: Proceedings of the 12th Annual Conference on Cyber and Information Security Research, Association for Computing Machinery, New York, NY, USA, CISRC '17 (2017a). https://doi.org/10.1145/3064814.3064831. https://doi.org/10.1145/3064814.3064831
62. Werner, G., Yang, S., McConky, K.: Time series forecasting of cyber attack intensity. In: Proceedings of the 12th Annual Conference on cyber and information security research, pp. 1–3 (2017)
63. Werner, G., Yang, S., McConky, K.: Leveraging intra-day temporal variations to predict daily cyberattack activity. In: 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, pp. 58–63 (2018)
64. Xu, M., Schweitzer, K.M., Bateman, R.M., Xu, S.: Modeling and predicting cyber hacking breaches. IEEE Trans. Inf. Forensics Secur. **13**(11), 2856–2871 (2018)
65. YILDIZ, K., BÜYÜKTANIR, B.: Comparison of lda, nmf and bertopic topic modeling techniques on amazon product review dataset: a case study. International Conference on Computing, Intelligence and Data Analytics (2023)
66. Zacharis, A., Patsakis, C.: AiCEF: an ai-assisted cyber exercise content generation framework using named entity recognition. Int. J. Inf. Secur. (2023). https://doi.org/10.1007/s10207-023-00693-z
67. Zacharis, A., Gavrila, R., Patsakis, C., Ikonomou, D.: Ai-assisted cyber security exercise content generation: Modeling a cyber conflict. In: 2023 15th International Conference on Cyber Conflict: Meeting Reality (CyCon), pp. 217–238 (2023). https://doi.org/10.23919/CyCon58705.2023.10181930