

Research of Networks of Cyber Security Subjects by Means of Generative Artificial Intelligence

Dmytro Lande^{1,2}, Anatolii Feher¹ and Leonard Strashnoy³

¹ National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" Kyiv, Ukraine

² Institute for Information Recording of NAS of Ukraine, Kyiv, Ukraine

³ University of California, Los Angeles (UCLA), USA

Abstract

The study of cyberwarfare as a concept is becoming an increasingly relevant area, using various approaches from identifying its components and dependencies to full-fledged conceptual forecasting using generative AI, which creates new opportunities for building capacious analytics. The article presents a methodology for ensuring the accuracy and completeness of NER processing of large data sets on the example of news clippings and articles on the topic of Israeli cyberspace attacks and explores the possibilities of using GPT for contextual prediction within the framework of peripheral associative series of the semantic network.

Keywords

cyberwarfare, semantic networks, prediction modeling, linguistics, N-GRAM, GTP, NER

1. Introduction

In the ever-changing cybersecurity landscape, the ability to effectively predict and counter cyber threats is critical. This study focuses on the application of semantic concept network prediction methods for object recognition and semantic pairing of related entities in the context of cyber warfare in Israel. It leverages advanced Natural Language Processing (NLP) tools to enhance predictive capabilities and countermeasures.


Named entity recognition (NER) offering a means to extract and classify key elements from unstructured text. In this study, these techniques are used to analyze a large number of news clippings and articles related to cyber warfare in Israel. The dataset, carefully selected from the open source space, provides a comprehensive view of all relevant cyber incidents.

The main goal of the work is to build and compare a predictive model capable of predicting pairs of inter-connected objects, entities in the semantic conceptual network of cyber actors. Using generative Artificial Intelligence (AI), in particular the Generative Pre-trained Transformer (GPT) model, the study aims to extract semantic relationships and identify key pairs of interconnected objects in the context of cyber warfare incidents.

An additional level of analysis is proposed to evaluate the predicted outcome of the extracted entity pairs by comparing it with the existing textual prediction method, thereby providing a scientific approach. The applied application of the described methods lies in the potential to enhance the capabilities of cyber incident analytics; this work aims to provide holistic information about patterns and trends within the topic of cybersecurity prediction.

ITS-2023: Information Technologies and Security, November 30, 2023, Kyiv, Ukraine

 dwlande@gmail.com (D. Lande); feher.anatolii@gmail.com (A. Feher); lstrashnoy@gmail.com (L. Strashnoy)

 0000-0003-3945-1178 (D. Lande); 0009-0007-4275-6905 (A. Feher); 0009-0008-5575-0286 (L. Strashnoy)



© 2023 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Methodology

Initial dataset consists of conducted news clippings and articles from open source resources via OSINT techniques, representing news about cyberattacks related to Israel. Dataset was prepared in the form of a text document for analysis with a capacity of 300 news clippings with an average length of 1178 words, on selected topics from English-language publications.

Irrelevant content such as date, author and links were initially filtered out, and the processed content was also divided into 6 parts for ease of further processing. Each of the created parts contained 50 news clippings, where each of the news was processed by generative AI individually with separate queries, without imprinting the previous query history, to ensure greater objectivity in the output data.

The processed raw data set from news clippings using NER techniques will be used to build a semantic conceptual network [1]. The approach involves iterative extraction of semantic relations from the text spaces, filtering them by the degree of interconnection and further its normalization. This multimodal approach using generative AI ensures fast processing, more accurate results, and flexibility in processing requests.

The dataset generated by the previous steps will represent the gold standard for building a semantic network, to determine the success of the predicted series created by N-gram autoregression and the GPT transformer relative to it. From the base dataset of 327 pairs capacity, where 31 pairs of low network connectivity with other nodes were cut off, which had the lowest number of edges and are less frequently repeated in the sample, where contextual network prediction will be aimed at reconstructing the missing peripheral nodes.

2.1. Original dataset processing

Named entity recognition is one of the fundamental components of NLP processing [2], which involves the identification and classification of named objects in text spaces into predefined categories. This process not only helps automate text processing itself, but also enriches its analytical potential, allowing for more efficient search and processing of information. By isolating entities such as people's names, company names, geographic coordinates, and other specific information, NER systems structure raw data. This is the technique used to extract semantic entities from the text space of news articles and form them into relevant interconnected pairs of entities that reflect the full semantic relationships of the clipping [3].

Recent advances in generative AI have changed the traditional approaches to NER, offering stable performance in a variety of linguistic landscapes and applications by generating queries or prompts. Based on the advantages of such generative models, we chose to build further processes on the GPT-4 transformer, which uses the Byte Pair Encoding (BPE) tokenization strategy, which understands the meaning in the text space and segments it into related subwords in several processing layers [4].

Such approach is especially effective for processing rare or complex lexemes, which increases the linguistic versatility of the model, where each token is transformed into a high-dimensional vector using an embedding matrix, which in turn encapsulates semantic and syntactic nuances [5]. This transformer architecture facilitates contextual embedding by dynamically adjusting to the space of neighboring tokens, this contextual awareness of the model is critical for interpreting complex patterns in texts, and allows for a fine-grained understanding of the relationships between extracted entities.

Despite its effectiveness, generative AI is not immune to inaccuracies caused by noise and hallucinations in the generated query responses, and such problems require manual validation and verification methods to ensure the integrity and correctness of the output data [6]. For this reason, proposed an integrated approach by duplicating each query of NER processing, such duplication of

the prompt, can be called iterative aggregation, which involves double processing of the input data to create multiple interpretations of this data.

The described strategy mimics a consultation with multiple virtual experts [7], where each of them offers their independent point of view on the outcome of the processed data set, thereby enriching the analytical depth and reducing the potential bias of only one answer. The integration of an iterative GPT query system into our methodology allows us to precisely control the number of query iterations, this software-driven approach not only ensures comprehensive data coverage but also increases the reliability of the results obtained through NER processing.

2.2. Interconnection percentage determining

To enhance result extraction with generative AI using Named Entity Recognition (NER), proposed an additional layer that assesses semantic relationships between identified object pairs in texts is crucial. This involves refining GPT models to evaluate the strength and nature of linguistic relations within single news articles.

The new layer incorporates both deterministic methods, such as word recurrence analysis, and probabilistic methods like cosine similarity between vector representations. These methods assess the frequency of term co-occurrence and the semantic proximity of entity pairs, respectively, providing insights into their contextual relationships.

A multimodal approach that blends these techniques ensures thorough data processing. By iteratively applying NER and analyzing interconnectedness through both methods, a comprehensive analysis is achieved. AI tools, customized through specific semantic analysis prompts, facilitate this integration. Each relationship level between entity pairs in news clippings is quantitatively assessed on a 0-100% scale, demonstrating the degree of linguistic relatedness.

Prompt: The initial extraction prompt:

Extract 25 pairs of semantically related and non-repeating entities, each entity should contain up to 4 words.

Each pair should be evaluated according to the criterion of recurrence and cosine similarity of the semantic relationship of entities represented as a percentage.

The results should be output in a strict CSV format in the form "entity1; entity2; Percentage of interconnectedness %".

Text for analysis: ...

This application, within a single logical query prompt, not only facilitates entity identification in textual spaces but also allows for a detailed analysis of their deep interrelationships while maintaining contextual integrity. This process is designed to enhance the accuracy of initial data by selectively filtering conceptual pairs, systematically excluding those with minimal entity interconnection. As a result, the dataset generated by the double query of 50 semantic pairs per news clipping is refined to a curated set of 20 pairs characterized by high relevance and accuracy potential for further work, proper visualizations shown in (Fig. 1).

Prompt: Secondary screening based on the principle of semantic connectivity.

Select the 20 most interconnected semantic pairs from the data set based on the percentage of linguistic connection between them.

Output the results in a strict CSV format as "entity1; entity2".

Pairs for analysis: ...

2.3. Linguistic pairs normalization

The further network representation of the already processed original data, using queries that consider both qualitative and quantitative aspects, poses significant challenges, especially due to the high linguistic discreteness in concept pairs. This is largely because of the presence of numerous synonyms or lexically similar terms among the extracted entities. Such variability in terminology introduces different linguistic categories of word cases, typically expressed through modifications like prefixes and suffixes. The linguistic heterogeneity generated by this input data often leads to distinctive patterns in the constructed semantic networks.

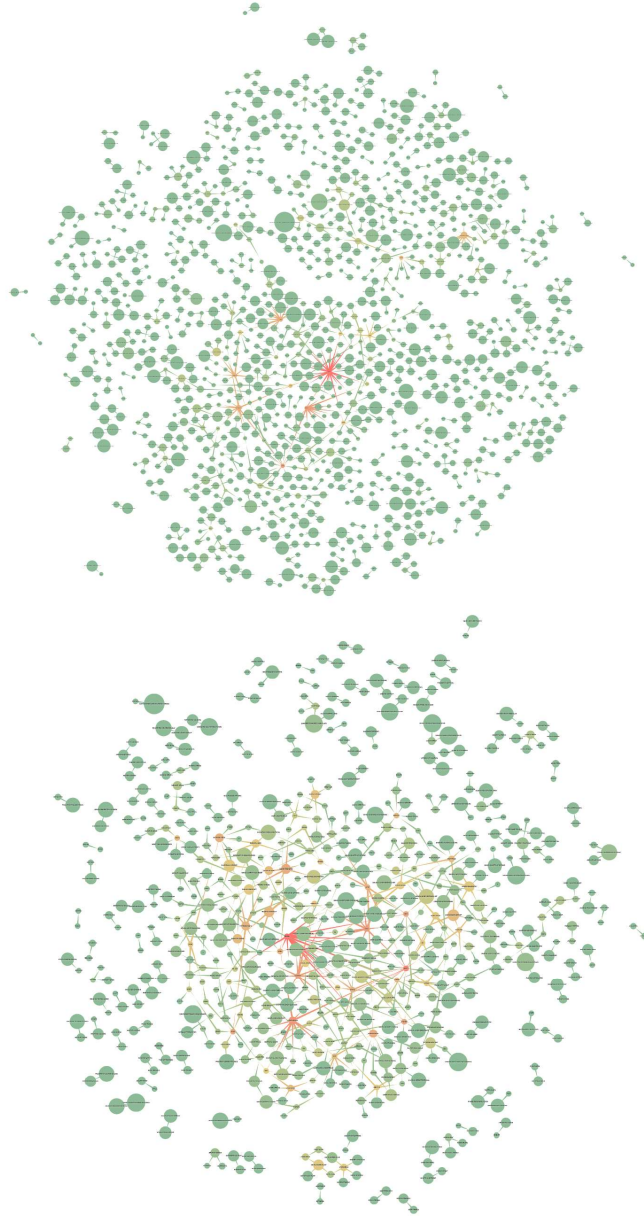


Figure 1: Upper image – generalized network of semantic pairs generated through first prompt, bottom image – represents filtered pairs with principle of semantic connectivity by second prompt

These variations complicate the identification of the most influential nodes, the assessment of network centrality, categorization application and clustering methodologies.

In practice, this issue is evident in variations of wording in entity pairs, such as “Israeli cyber incident” versus “cyber incident in Israel”. To mitigate these differences, lexical normalization methods are employed to enhance the structural integrity of the semantic network. Generative AI

excels at normalization tasks due to its deep contextual understanding, which improves the efficiency and accuracy of identifying relationships between terms. Using GPT models automates the identification of linguistically similar phrases and suggests normalized forms, which can be automatically incorporated to update the dataset for network building, standardizing all variants to a single form recommended by the AI.

To normalize similar entities efficiently, the proposed method uses language clustering, grouping closely related entities iteratively until a similarity threshold is met. This process includes deduplication within each cluster to ensure homogeneity, supported by a specialized prompt in data processing to enhance lexeme consistency and accuracy.

Prompt: Further normalization of preformed semantic pairs:

For each pair of provided entities, simplify them to 2-3 words, structure and normalize all synonyms of entities and linguistic duplicates in the data.

Save the results in a strict CSV format as "entity1; entity2".

Pairs for analysis: ...

Thus, the described step-by-step multimodal approach consisting of three hierarchical queries iteratively extracts semantic relations from the news space, evaluates the formed pairs of such relations by the level of their interconnectedness, and then filters the processed dataset, leaving the most relevant pairs of entities, which normalizes the tokens in the pairs within the newly formed data set. The result of this process is a holistic representation of the base sample that fully illustrates the entire semantic load of 300 news articles shown in (Fig. 2).

2.4. N-gram forecasting

N-gram models are foundational tools in computational linguistics and statistical natural language processing, relying on the Markov assumption that the probability of a word depends only on its preceding $n-1$ words. By analyzing previous elements in a sequence, these models provide a probabilistic approximation of linguistic structure, capturing local context up to $n - 1$ elements. While simple and widely applied across tasks such as text prediction, speech recognition, and music generation [8], n-gram models are constrained by their surface-level statistical approach and the high computational costs associated with larger n values, leading to the "curse of dimensionality".

To mitigate these limitations and improve predictive accuracy, specific adaptations have been implemented. One notable adaptation is the bigram model, which reduces computational demand and scarcity issues by predicting the next word based on its immediate predecessor only, enhancing efficiency in scenarios that require the prediction of entity pairs.

Building on these foundational adaptations, the application of frequency weights in n-gram models introduced a refined approach to sequence prediction. Instead of selecting the next term arbitrarily, applied method of frequency-based weighting design to prioritize transitions that occur more frequently, creating a probabilistic model that more closely mirrors real-world language use. This adaptation helped not only addresses the model's tendency to falter with rare sequence combinations but also enhances its ability to capture and predict common linguistic patterns in pairs more effectively.

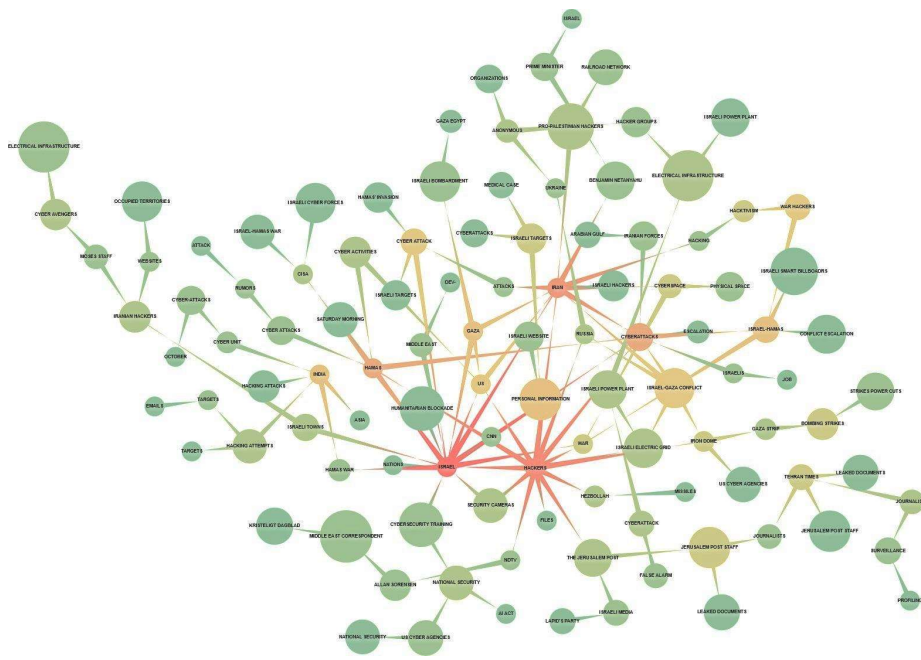


Figure 2: Normalized conceptual semantic network that represents inner semantics of Israel cyberwarfare news

Furthermore, the backoff strategy complements the frequency weights by offering a solution to situations where the current term provides limited or no onward transitions. Under this strategy, if a higher-order n -gram does not yield a prediction due to data scarcity, the model "backs off" to a lower-order n -gram, utilizing less specific contexts. This method can also involve random sampling from all available terms when necessary, thereby maintaining predictive continuity even in the absence of strong statistical evidence. Applied strategy ensures robustness and adaptability in dynamic linguistic environments.

Described enhancements as frequency weights and backoff are crucial for extending the utility of n -gram models beyond their traditional limitations [9]. By integrating these strategies, n -gram model achieved a more nuanced balance between performance and computational efficiency, maintaining its status as critical baselines in linguistic performance and facilitating comparative analyses against more complex models like neural networks or hidden Markov models. This made it possible to efficiently and accurately process a dataset of 296 pairs, where the discarded 31 pairs representing the weakest links in the network were used to compare the effectiveness of the improved n -gram prediction method.

2.5. GPT forecasting

The integration of generative AI into predictive analysis has attracted considerable interest due to the complexity and linguistic specificity inherent in their transformational nature, such models excel at processing large text sequences, thereby capturing complex dependencies and relationships between words over long distances. The key aspect of these models is their sophisticated attention mechanism: unlike traditional models, GPT uses a dynamic form of attention that changes the weight of different segments of the input sequence during the prediction task [10]. This selective attention contributes to the deep contextual understanding required to make accurate predictions.

The architectural basis of GPT models is built on a multi-layer autoregressive framework, which typically consists of hundreds of transformation layers. Each such layer is equipped with several self-monitoring heads that work independently to dissect and analyze different segments of the

input data. These heads help to detect various linguistic patterns and syntactic dependencies, enriching the overall contextual accuracy of the model.

This multi-faceted capability is the result of a complex training process for such models, which consists of two main stages, where the model first undergoes pre-training, where it learns general linguistic patterns from a large corpus of data. After that, it moves to the fine-tuning stage, where it adapts to specific forecasting tasks by adjusting its parameters to the nuances of the specific data. The described two-level training structure allows GPT models to generate highly relevant and contextually consistent forecasts.

However, despite their robust capabilities, GPT models are not without limitations. One of the main problems is the limited window of context, which limits the model's ability to process very long sequences in a single operation. This limitation often results in the need to segment large datasets into smaller, manageable chunks for processing, potentially leading to fragmentation of contextual continuity. In addition, these models are prone to producing hallucinations, generating a plausible but contextually irrelevant or factually incorrect result, therefore, decided to proceed with GPT-4 rather than GPT-3.5 [11].

Due to this problem, the resulting predicted set of semantic pairs of entities was carefully processed in manual verification mode after several generations of answers to the generated query for filling in peripheral network nodes to ensure the appropriate relevance and accuracy of the output data.

Prompt: Contextual prediction generation request:

Please leverage your extensive autoregressive predictive abilities to forecast the next 31 pairs from the associative semantic network dataset. Focus on understanding the context it relations.

Dataset list used for examination: ...

3. Results and Discussion

The final dataset for building the network was formed by described multimodal approach using generative AI, based on the approaches of iterative extraction of semantic units with the identification and subsequent screening by the criterion of their level of interconnectedness, and subsequent normalization of the discreteness of entities in pairs, and was used as a gold standard for visualization, comparison, and determination of the accuracy of the generated contextual predictions. Such concluded semantic network dataset that contain 327 interconnected pairs generated all along with GPT-4 presented in example.

Example: Part of semantic network dataset:

US; IRAN
ARABIAN GULF; IRANIAN FORCES
MISSILES; HEZBOLLAH
IRANIAN FORCES; CYBERATTACKS
GAZA; IRAN
SECURITY CAMERAS; HACKERS
ISRAEL; CNN
MIDDLE EAST; ISRAEL
TEL AVIV; AARON KROLIK
ISRAEL; HAMAS WAR
ANONYMOUS; HACKER GROUPS
SCAM EMAILS; DONATIONS
ISRAELI SECURITY; IDF
CYBERSECURITY TRAINING; ISRAEL

HACKER GROUPS; ISRAELI SECURITY ESTABLISHMENT
 CYBERATTACK; ISRAELI POWER PLANT
 HACKERS; ISRAELI ELECTRIC GRID
 HUMANITARIAN BLOCKADE; ISRAEL
 RUSSIA; UKRAINE
 X; HAMAS ACCOUNTS
 WAR HACKERS; HACKTIVISM
 CANADIAN WATCHDOG; SPYWARE
 SPYWARE; NSO GROUP
 MAHMOUD ABBAS; PALESTINIAN AUTHORITY
 JEWS; USURY
 ISRAELI POWER PLANT; ELECTRICAL INFRASTRUCTURE
 RUSSIAN JOURNALIST; PEGASUS SPYWARE; NSO GROUP
 PRO-PALESTINIAN HACKERS; RAILROAD NETWORK
 TEHRAN TIMES; LEAKED DOCUMENTS
 ...

With the described methodology, directed semantic networks were crafted for both the basic dataset and the augmented predicted series of semantic entity pairs. The visualizations present extracted concepts and their interconnectedness pertinent to the cyber war in Israel, its actors, and valuable elements, where the size of the visualized entity nodes correlates with the length of the entity lexeme. Connectivity among the entity nodes is described with a color-coded scheme where green represents low-level nodes, yellow denotes medium-level nodes, and red highlights the most connected nodes, chosen employed as the usual color scheme to signify the importance and connectivity of the nodes, while the edges illustrate the degree of incoming and outgoing connections for each entity node.

To assess the accuracy of the predicted semantic pairs of entities, the F1 score was used as the evaluation metric, this score is particularly valuable for studies because it helps ensure that both the model's precision and its ability to capture relevant instances (recall) are considered. The F1 score that presented in equation 1 is beneficial when the data might have uneven class distributions and both false positives and false negatives are important to measure. Precision is defined as the ratio of correct positive predictions (true positives) to all positive predictions made (the sum of true positives and false positives). It shows how accurate the predictions are. Recall, on the other hand, measures the ratio of correct positive predictions to the total number of actual positive cases in the data (the sum of true positives and false negatives). It indicates how well the model identifies all relevant instances.

$$F1 = 2 \times \frac{\textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}} \quad (1)$$

Based on the formula calculation means that a high F1 score can only be achieved when both precision and recall are high, this encourages a balance in the model, ensuring it does not favor precision over recall, or vice versa. For this study, using the F1 score allowed for a thorough evaluation of how well the N-gram and generative AI model could predict and link relevant semantic entities, this approach helps confirm the model's predictive reliability.

Furthermore, the evaluation of how accurately models predicted each pair of entities involved using cosine similarity, the method that measures how closely two linguistic vectors are aligned, which the study, represents the predicted and actual entity pairs comparison. Each entity pair was turned into a vector, and cosine similarity was calculated by measuring the angle between these vectors. It unveils how similar the predicted pairs are to the actual pairs, analysis utilized two key metrics: the calculation of true positives, false positives, and false negatives; and the measurement

of cosine similarity, as delineated in Equation 2, to quantify the alignment between the predicted and actual data pairs.

$$Similarity = \frac{A \cdot B}{\|A\| \|B\|} \quad (2)$$

Initially, the analysis calculated the number of true positives, which are pairs that the model predicted correctly according to the base data, it also counted false positives, which are incorrect predictions not found in the base data, and false negatives, which are correct pairs missed by the model. The average similarity of the true positives was calculated and is displayed in Table 1. These numbers help determine the average similarity level between predicted and actual pairs, that lead to more deep comparison of employed methods.

Table 1
Average similarity behind correct predictions

	N-gram	GPT model
Average Similarity	0.776	0.619

A threshold of 0.5 for cosine similarity was chosen as the standard cutoff point, predictions with a lower angle of vector that was above the defined threshold were considered true positives, meaning they matched well with the actual data, showing that the model predicted prereferral nodes and their connections correctly. Predictions with a higher angle of vector that was below the defined threshold were marked as false positives, indicating errors in the model's predictions. Additionally, any correct pairs not predicted by the model and scoring above this threshold were noted as false negatives, pointing out where the model missed predicting true entity relationships.

The results, including each pair's cosine similarity score and its classification as either a true positive, false positive, or false negative, are summarized in Table

This table helps to easily see how well the N-gram and GPT node predictions align with the actual cut prereferral data and is key for assessing the model's effectiveness in identifying the relationships within the thematic semantic networks.

Table 2
Results from cosine similarity metric

	N-gram	GPT model
TP	17	19
FP	14	12
FN	31	31

Predicting nodes with low connectivity or peripheral positions within the network poses a significant challenge, these nodes are typically less central and have fewer connections, making accurate predictions difficult, often due to limited data and weaker signals in the relationships between terms. Nevertheless, the F1 Score represents moderate, but still valuable results as shown in Table 3.

Table 3

Comparison of N-gram and GPT model predictions accuracy

	N-gram	GPT model
Precision	0.548	0.613
Recall	0.354	0.381
F1 Score	0.430	0.469

The precision of the GPT model is quantified at 61.3%, achieved statistic reveals that when the model classifies a node as a peripheral node with low connectivity, it accurately does so approximately 61.3% of the time. Given the minimal adverse consequences of falsely identifying a node as peripheral in the broader scheme of semantic network analysis, this level of precision is deemed satisfactory, however, ongoing efforts to refine the model’s accuracy are crucial, particularly in scenarios where higher stakes decisions depend on the precise identification of node characteristics.

The recall rate of the model is calculated to be 38.1%, which indicates that the model effectively identifies 38.1% of all true peripheral nodes within the network. Although this recall rate might initially appear low, it reflects the inherent difficulty in detecting peripheral nodes, which typically exhibit fewer connections and less distinct features compared to more central nodes. Strategies to improve this metric are vital, particularly for applications where comprehensive detection of peripheral nodes is essential to maintaining network integrity and functionality.

The F1 Score, a harmonic mean of precision and recall, is reported at 43% for the N-gram model and 46.9% for the GPT model, calculated scores highlight an opportunity for enhancing the model’s performance. The moderate F1 Score suggests that balancing the optimization of both precision and recall could significantly boost the model’s effectiveness in practical prediction scenarios.

Regarding the similarity scores, the GPT model capabilities seem to vary more or cover a broader linguistic range than those from the N-gram model, which resulted in lower average similarity scores, with predictions that slightly exceed standard deviations. This suggests and can be interpreted that the GPT model might be better at generalizing and capturing a wider range of relevant features, which tends to be more accurate than depicted with an F1 score as well as more novelty predictions, despite less overall similarity in vector orientation. To further appreciate the variety and novelty in the predictions made by the GPT model, additional metrics and analyses such as Entropy, Distinct-1, and Distinct-2 metrics were used to look at the diversity and deeply understand predictive capabilities within the predicted semantic pairs in the network.

Entropy, a concept that evaluates the diversity of information, borrowed from information theory, which quantifies the randomness or unpredictability in the researched dataset, calculates with an equation 3 and leads to assess whether the models tend to generate a broad spectrum of predictions or if they are biased towards repeating certain pairs more frequently. A higher entropy value indicates a more diverse set of predictions, suggesting that the models are capable of capturing a wider variety of relationships within the semantic network.

$$H(X) = -\sum_{i=1}^n p(x_i) \log p(x_i) \quad (3)$$

To enrich methods analysis evaluated the uniqueness of the model predictions with employed Distinct-1 and Distinct-2 metrics, which can unveil insightful trends about the performance capabilities. Distinct-1 that describes with an equation 4 measures the number of unique single words (unigrams) as a proportion of the total number of words in the predicted pairs, based on outputs from N-gram and GPT models, providing insight into the lexical variety of the prediction pairs. Distinct-2 shown with an equation 5, on the other hand, calculates the proportion of unique

consecutive word pairs (bigrams) to the total number of bigrams, also based on the predictions from these models, such metric is particularly useful for assessing the diversity of two- word combinations in the models' output. High scores in Distinct-1 and Distinct-2 indicate a high level of novelty and variation in the text generated by the models, reflecting their ability to create a range of different and contextually appropriate semantic pairs.

$$Distinct - 1 = \frac{Number\ of\ unique\ unigrams}{Total\ number\ of\ unigrams} \quad (4)$$

$$Distinct - 2 = \frac{Number\ of\ unique\ bigrams}{Total\ number\ of\ bigrams} \quad (5)$$

The inclusion of entropy, Distinct-1, and Distinct-2 metrics, based on N-gram and GPT predictions, enhances the analysis by providing a holistic view of model performance, not just in terms of accuracy but also in predicting diverse and novel semantic pairs, as detailed in Table 4. These metrics ensure that the models not only predict accurately but also produce varied and contextually rich semantic pairs.

A closer examination of the results from the entropy, Distinct-1, and Distinct-2 metrics reveals insightful trends in the performance of contextual predictive capabilities of the N-gram and GPT models. Both models demonstrate high entropy values, indicative of their ability to produce a varied and unpredictable array of outputs. The GPT model, with an entropy score of 4.87, exhibits greater unpredictability compared to the N-gram model's score of 4.70. This suggests that the GPT model might be better suited for tasks requiring the generation of novel and varied outputs without a fixed pattern.

Table 4

Comparison of N-gram and GPT model predictions novelty and performance

	N-gram	GPT model
Entropy	4.70	4.87
Distinct-1	0.73	0.66
Distinct-2	0.91	0.93

In terms of Distinct-1, which measures the uniqueness of single words within the outputs, the N-gram model scores higher (0.73) than the GPT model (0.66), such results indicate that the N-gram model slightly utilizes a broader vocabulary in its network predictions. Such a characteristic can be particularly advantageous in applications where lexical diversity is critical, such as in semantic analysis or content generation where a wider range of vocabulary could enhance the comprehensive- ness and depth of the analysis.

For Distinct-2, which focuses on the uniqueness of bigrams, both models show high values, with the GPT model scoring slightly higher (0.93) than the N-gram model (0.91), such high performance in generating unique bi-grams demonstrates both models effectiveness in constructing diverse and contextually appropriate two-word combinations, that especially fit particular interconnected semantic pair predictions. The slight edge of the GPT model in this metric could indicate its superior capability in capturing and generating more complex, contextually nuanced relationships between words.

These findings highlight the strengths and potential uses of each model for tasks like predicting peripheral nodes in networks. The GPT model excels in generating innovative and complex outputs, suitable for dynamic and creative applications, while the N-gram model is valued for its diverse vocabulary, critical for in-depth semantic analyses. Additional visual analysis of network

representations and node predictions (Fig. 3) shows the GPT model's capacity to identify novel, out-of-system nodes, highlighted in pink.

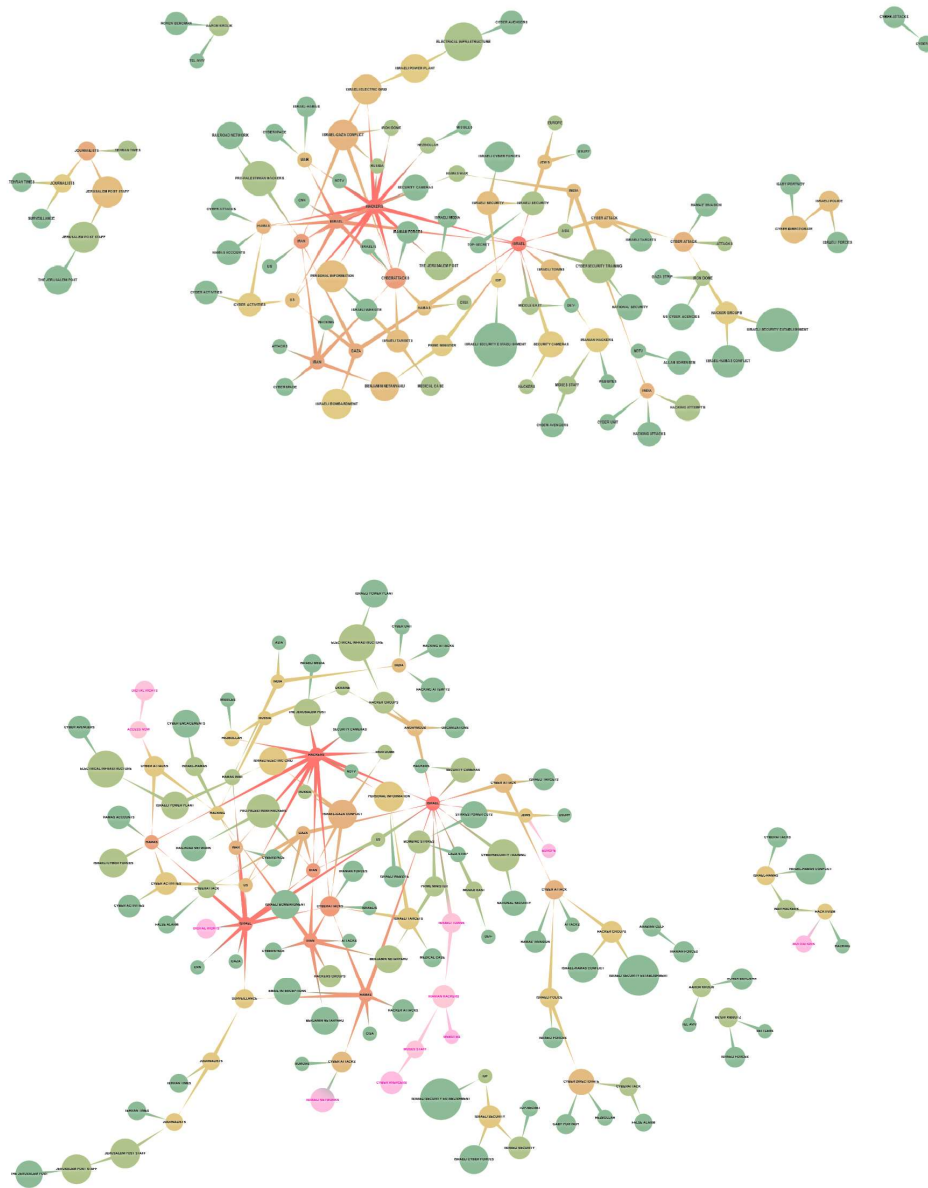


Figure 3: Upper image – N-gram contextual network prediction, bottom image – GPT contextual network prediction. Pink – represents new branches of nodes that were predicted under cutted by 31 pairs base semantic network

Despite some inefficiencies, it's important to consider that GPT models have greater potential due to their rapid development and the ability to go beyond standard regression algorithms by offering new sequences as predicted series. This capability introduces some instability and uncertainty [12], but it also holds practical applications in the real world. The issue of extrapolation and going beyond the data for processing or learning is considered risky due to assumptions that may not be true, yet it is precisely this ability to innovate beyond conventional limits that makes these models particularly valuable in dynamic fields such as cybersecurity.

4. Conclusion

The conducted research has emphasized the crucial role of enhancing analytics in cyberwarfare and incident analysis through the utilization of semantic network contextual prediction measures. The effectiveness of Named Entity Recognition (NER) methods combined with advanced generative AI models like GPT-4 has been demonstrated. While the prediction capabilities of the GPT model exhibited a precision rate of 61.3%, efficiently identifying peripheral nodes with low connectivity, its recall rate of 38.0% highlights some limitations in capturing nodes with sparse connections. In contrast, predictions of more central nodes in a semantic network, typically involve nodes with stronger and more numerous connections and tend to be simpler due to the more evident semantic relationships.

Further comparison with N-gram models has illuminated the potential of GPT models for sequence prediction enhancements, such models adaptability in generating new sequences is beneficial, enhancing both precision and recall. However, the inherent risk of generating inaccurate predictions underscores the importance of their careful application. Tailoring fine-tuning techniques to the specific requirements of cyberwarfare analytics is essential to mitigate the incidence of erroneous predictions and to ensure their relevance and reliability.

Additionally, the inclusion of entropy, Distinct-1, and Distinct-2 metrics in this study has provided a deeper insight into the models operational capabilities, both models demonstrated high entropy values, indicating their ability to generate varied and unpredictable outputs, which is essential for dynamic cyber environments. The N-gram model's higher Distinct-1 score suggests a broader vocabulary, enhancing comprehensive semantic analyses, while the GPT model's slightly higher Distinct-2 score indicates its superior capability in generating complex, contextually appropriate bi-grams. These findings underline each model's strengths and shed light on their suitability for different aspects of cybersecurity applications.

Conclusively, it can be affirmed that both studied methods operate effectively, supporting the potential for predictions within network structures to have practical applications in predicting network dynamics and enriching structured networks with external factors. These factors are crucial for analytical work and strategy development in cyberwarfare. Utilizing methods to reduce prediction errors, enhancing model relevance through fine-tuning, and employing multimodal prediction methods to improve output accuracy are pivotal, such analytics provide a robust basis for drawing comprehensive conclusions about the choice of effective strategies and countermeasures in the cyber warfare space.

References

- [1] Lande D., Strashnoy L. GPT Semantic Networking: A Dream of the Semantic Web - The Time is Now. — ISBN 978-966-2344-94-3. — 2023. — 168 p.
- [2] Indurkha N., Damerau F. J. Handbook of natural language processing. — CRC Press, 2010. — 2010. — 676 p.
- [3] Feher A., Lande D. Defined AI semantic networking in cybersecurity // Intelligent Solutions-S: Proceedings of the International Symposium, Kyiv-Uzhorod, Ukraine. Kyiv: Publishing House "Caravela", ISBN 978-966-801-916-6. — 2023. — P. 21– 22.
- [4] Martin Berglund B. v. d. M. Formalizing BPE Tokenization // 13th International Workshop on Non- Classical Models of Automata and Applications. — 2023. — P. 16–27. — DOI: 10.4204/EPTCS.388.4.
- [5] Jianpeng Cheng L. D., Lapata M. Longshort-term memory-networks for machine reading // Conference on Empirical Methods in Natural Language Processing. — 2016. — P. 551–561. — DOI: 10.48 550/arXiv.1601.06733.

- [6] Anne-Dominique Salamin D. R., Rueger D. ChatGPT, an excellent liar: How conversational agent hallucinations impact learning and teaching // 7th International Conference on Teaching, Learning and Education. — 2023. — P. 551–561.
- [7] Lande D., Feher A., Strashnoy L. Cybersecurity in AI-Driven Casual Network Formation // Theoretical and Applied Cybersecurity. — 2023. — Vol. 5–Issue 2. — P. 105–113. — DOI: 10.20535/tacs.2664-29132023.2.287139.
- [8] Sridharan A. Music Similarity Estimation. — 2018. — DOI: 10.31979/etd.8nz2-b9ya.
- [9] Bo-June (Paul) Hsu J. G. N-gram Weighting: Reducing Training Data Mismatch in Cross-Domain Language Model Estimation // Conference on Empirical Methods in Natural Language Processing. — 2008. — P. 829–838.
- [10] Benyamin Ghogh A. G. Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey. — 2020. — DOI: 10.31219/osf.io/m6gcn.
- [11] Koubaa A. GPT-4 vs. GPT-3.5: A Concise Showdown. — 2023. — DOI: 10.20944/preprints202303.0422.v1.
- [12] Jeff Mitchell Pasquale Minervini P. S., Riedel S. Extrapolation in NLP // Proceedings of the Workshop on Generalization in the Age of Deep Learning. — 2018. — P. 28–33.