Bilingual Text Representation Using Knowledge Graphs in the Cybersecurity Domain

Bohdan Pavlyshenko[1,*,†] and Mykola Stasiuk[1,†]

[1] *Ivan Franko National University of Lviv, Drahomanova 50, 79000 Lviv, Ukraine*

Abstract

This paper presents a bilingual framework for constructing and analyzing knowledge graphs derived from English and Ukrainian cybersecurity reports. The framework integrates large language models, multilingual sentence embeddings, and analytics based on graphs to extract, structure, and connect entities and relations across languages. Using LaBSE embeddings and HDBSCAN clustering, the system identifies semantically coherent entity groups without predefined taxonomies. LLMs further enhance interpretability through automated cluster labeling, while a cross-lingual alignment module links equivalent clusters between languages based on semantic and lexical similarity. The resulting bilingual knowledge graph captures shared and language-specific cybersecurity concepts, enabling joint reasoning across corpora. Finally, association rule mining uncovers recurrent co-occurrence patterns between entities, validating semantic consistency across languages. The proposed approach demonstrates robustness and scalability for multilingual knowledge discovery in the cybersecurity domain, providing a foundation for unified, cross-lingual threat intelligence analysis.

Keywords

knowledge graphs, large language models, bilingual analysis, multilingual embeddings

1. Introduction

The modern digital landscape is defined by an ever-increasing volume and velocity of cybersecurity data. From threat intelligence reports and vulnerability databases to security incident logs and regulatory compliance documents, this data is often disparate, unstructured, and siloed. This growing complexity makes it extremely challenging for human analysts and automated systems alike to gain a coherent, comprehensive understanding of the evolving threat environment. Consequently, there is a need for sophisticated methods of structured knowledge representation that can integrate, contextualize, and accelerate the analysis of this complex information.

Knowledge Graphs (KGs) have emerged as a powerful solution to this issue. By representing entities and their semantic relationships in a graph structure, KGs transform raw, fragmented data into actionable, interconnected intelligence. This allows for superior pattern recognition, root cause analysis, and proactive defense strategy formulation.

However, a significant, often overlooked challenge in threat intelligence is its multilingual nature. As geopolitical conflicts and global supply chains intensify, threat intelligence reports, security advisories, and even malware communication often originate in or must be translated for multiple linguistic contexts. Focus of this work is to addresses the need to handle data in both English and Ukrainian. Successfully integrating and analyze data from both language streams that can be helpful for security analysts operating in multinational or conflict-adjacent environments, yet maintaining conceptual consistency across these linguistic divides is technically demanding.

This paper addresses the gap by outlining a methodology for constructing bilingual knowledge graphs in the cybersecurity domain. Primary objective is to leverage the advanced capabilities of Large Language Models (LLMs) to automatically extract, structure, and populate distinct knowledge graphs for English and Ukrainian text sources. The final goal is not merely to create two separate KGs, but to develop and evaluate a method for aligning their thematic structures. This

[1] *SMICS'25: Workshop on Cryptology and Data Security, October 16-18, 2025, Lviv, Ukraine*

[*] Corresponding authors.

[†] These authors contributed equally.

bohdan.pavlyshenko@lnu.edu.ua (B. Pavlyshenko); mykola.stasiuk@lnu.edu.ua (M. Stasiuk)

0000-0001-9515-3488 (B. Pavlyshenko); 0009-0006-5256-2103 (M. Stasiuk)

alignment will allow scientists to seamlessly cross-reference information, ensuring that a threat identified in one language is semantically linked to and understood in the other, thereby creating a unified, resilient, and comprehensive bilingual cybersecurity knowledge base.

2. Related Work

Research into Cybersecurity Knowledge Graphs (CSKGs) has a robust foundation built upon established standardization efforts. Frameworks like the Structured Threat Information eXpression [1], often disseminated via the Trusted Automated Exchange of Indicator Information [2] protocol, provide a standardized vocabulary and structure for describing threat intelligence, including indicators of compromise, threat actors, and attack patterns. Furthermore, the MITRE ATT&CK [3] framework has become the de facto industry standard for modeling adversary tactics and techniques. These frameworks serve as foundational ontologies for many CSKGs, providing a common ground for representation. However, while these standards are excellent for domain modeling, they often fall short in providing scalable, automated methods for multilingual knowledge extraction from unstructured text sources, which is a key focus of this study. Most existing CSKGs are predominantly built upon or optimized for English-language data.

The emergence of LLMs has revolutionized the field of Information Extraction (IE), offering powerful capabilities for transforming unstructured text into structured data [4, 5, 6]. LLMs are being leveraged for entity extraction and text classification within the different domains, like medical, scientific, cybersecurity and other [7, 8, 9]. These models demonstrate superior performance compared to traditional rule-based or statistical methods, especially in handling complex, nuanced, or evolving threat language. Recent studies have focused on fine-tuning specialized LLMs for cybersecurity tasks, effectively automating the first step of KG construction [10, 11]. This work specifically utilizes the zero-shot and few-shot learning capabilities of LLMs [12] to build comparable extraction pipelines for two distinct languages without requiring extensive, manually-labeled parallel corpora.

The proposed alignment methodology stands on the shoulders of significant work in bilingual and cross-lingual Natural Language Processing (NLP). Models like Multilingual BERT [13] and Language-Agnostic BERT Sentence Embeddings (LaBSE) [14] have demonstrated the capability to project text from different languages into a shared vector space, allowing for cross-lingual transfer learning and similarity search. These models are typically trained on massive amounts of diverse multilingual text, learning to map conceptually similar sentences across languages closer together in the embedding space. They have been successfully applied in general domain tasks such as cross-lingual document classification [15, 16, 17] and information retrieval [18, 19]. These advancements were adapted to the highly specialized and constantly evolving terminology of cybersecurity, using their vector space alignment properties to address the challenge of thematic alignment between English and Ukrainian knowledge graphs.

While the aforementioned approaches have set high standards, significant limitations persist in the context of multilingual cybersecurity. STIX/TAXII and ATT&CK lack native mechanisms for automated, large-scale, and semantic cross-lingual linking [20]. Furthermore, the reliance of current LLM-based extraction on high-resource languages often results in suboptimal performance for less-resourced languages [21], particularly given the domain-specific jargon [22]. Most importantly, existing cross-lingual NLP models are typically evaluated on sentence or document-level alignment, but they do not provide a direct, actionable mechanism for aligning the graph structures and semantic entity relationships themselves [23]. This research directly addresses these limitations by developing a framework for not just building, but semantically aligning two separately constructed knowledge graphs to achieve bilingual knowledge integration.

3. Data Collection

The core of this research relies on a bilingual corpus of real-world cybersecurity threat intelligence sourced from two highly reputable and distinct organizations: Computer Emergency Response Team of Ukraine (CERT-UA) [24] and Cisco Talos Intelligence Group (TALOS) [25]. The selection of these sources was deliberate, ensuring both linguistic diversity and thematic relevance to contemporary cyber threats, particularly those impacting the Ukrainian digital landscape.

CERT-UA serves as the primary source for the Ukrainian-language corpus. As an entity responsible for responding to cyber incidents against critical infrastructure and government bodies in Ukraine, its official reports provide authentic, high-impact intelligence characterized by unique Ukrainian terminology and a direct focus on threats relevant to the national context.

TALOS Intelligence was selected as the source for the English-language corpus. TALOS is a globally recognized threat intelligence organization that provides comprehensive, technical analysis of malware, vulnerabilities, and emerging threat actors. This ensures the English corpus maintains the high technical complexity and broad international scope necessary for a robust cybersecurity knowledge graph.

To systematically build the corpus, an automated web scraping methodology was employed. This approach was essential due to the dynamic nature of both websites, which frequently update their content.

1. The scraping process utilized a combination of Playwright [26] and BeautifulSoup [27]. This combination allowed for the reliable extraction of article text from within the structures of both CERT-UA and TALOS.
2. Once the dynamic content was loaded, the relevant HTML elements containing the main article body were parsed. The text content of every individual article was then scraped directly and saved as a separate .txt file.
3. All collected text files were subsequently collated and processed into a .csv file for each language. Each row in the final CSV dataset contained a minimum of the article's Title, URL and the complete extracted text body, facilitating streamlined pre-processing for the subsequent LLM-based extraction phase.

The constructed dataset was designed as a proof-of-concept to validate the proposed LLM-based KG construction and alignment methodology. To this end, a balanced corpus of 200 documents per language was collected, resulting in a total dataset size of 400 cybersecurity intelligence articles. While this number is relatively small compared to massive, single-language corpora, it was deemed sufficient for:

- Demonstrating the feasibility of automated, high-quality knowledge extraction in a complex domain using LLMs.
- Testing the effectiveness of the thematic alignment process between the two distinct linguistic knowledge graphs.
- Providing a highly targeted, current, and domain-specific bilingual resource, thereby ensuring the generated knowledge graphs are immediately relevant to real-world threat analysis.

4. Methodology

4.1. Entity and Relation Extraction

Entity and relation extraction form the foundational layer for constructing the proposed bilingual cybersecurity KG. The primary objective of this stage was to systematically identify relevant named entities and to infer semantic relations between them. To ensure cross-lingual consistency,

high precision, and robustness against the complex jargon of threat intelligence, a hybrid extraction pipeline was employed for both the English and Ukrainian corpora. This approach combined the contextual understanding capabilities of a LLM with the precision of rule-based and lexical components specialized for the cybersecurity domain.

4.1.1. English Corpus Processing

The processing of the English-language corpus utilized a two-pronged strategy:

- LLM-Based Core Extraction;
- Rule-Based and Lexical Augmentation.

Core entity and relation extraction was performed using GPT-5 [28], which provided state-of-the-art contextual understanding and robust semantic reasoning. The model was instructed to analyze each article at the sentence level and produce entity-relation pairs directly in a structured JSON format, following a predefined schema. Although exact reproduction of GPT-5 outputs is not possible due to the proprietary and non-deterministic nature of large instruction-tuned models, the extraction process remains methodologically transparent. The overall design, including the entity categories, relation types, and general processing flow, can be replicated using comparable open-weight instruction-tuned models. While surface-level text may vary between models, the resulting graph structure and information coverage remain consistent, ensuring methodological transparency and reproducibility of the overall framework, even if model-level outputs remain stochastic.

The schema ensured the output was immediately machine-readable, classifying each extracted entity into one of the standardized categories necessary for threat intelligence modeling: MALWARE, CVE, VULNERABILITY, ORG, PRODUCT, INDICATOR, ACTOR, LOCATION, or OTHER.

To ensure high recall and precision for structured artifacts, the LLM output was complemented by several deterministic layers:

- Regular Expressions: Explicit regular expressions were used to identify canonical structured artifacts, including:
    a. CVE patterns (e.g., CVE-d4-d4,7);
    b. IPv4 addresses (e.g., $\text{\b\d{1,3}(?:\.\d{1,3}){3}\b}$);
    c. URLs (e.g., $\text{https?://[^\s]+}$);
    d. Hashes (e.g., [A-Fa-f0-9]32,64).
- Lexical Dictionaries: Dedicated dictionaries were used to match known cybersecurity entities, such as common malware names and authoritative organizations;
- Contextual Pattern Triggers: Deterministic templates were used to infer relations based on linguistic cues, enhancing the confidence of relations extracted by the LLM. Examples included mapping "reported by" to reports, "used in attacks" to uses, and "exploited via" to exploits.

Each extracted entity and relation record was assigned a confidence score, reflecting the level of agreement between the extraction mechanisms:

- High Confidence: Confirmed by both the LLM and either a regex match or a lexical dictionary trigger;
- Medium Confidence: Detected by only one primary mechanism;
- Low Confidence: Uncertain or syntactically ambiguous cases, flagged for potential manual review.

4.1.2. Ukrainian Corpus Processing

The processing of the Ukrainian corpus maintained the parallel structure and entity taxonomy to ensure the resulting KG was fully interoperable with the English graph.

The Ukrainian corpus was also processed using GPT-5's native multilingual comprehension. This design choice allowed for direct entity recognition and relation extraction from the Ukrainian text without requiring an intermediate, lossy translation step. This approach leveraged the LLM's capability to understand domain-specific context in a less-resourced language while maintaining the same structured JSON output schema as the English pipeline.

The LLM output was refined and augmented using language-specific detectors to address the unique linguistic characteristics of Ukrainian cybersecurity reports:

- Deterministic Artifacts: The same regular expressions were used to identify canonical artifacts often present in both languages;
- Organization Lexicons: Specific organization lexicons were compiled to include Ukrainian government and state security entities alongside international names;
- Malware Name Detection: Specialized logic was included to detect English-named malware families frequently embedded directly within the surrounding Ukrainian text;
- Linguistic Relation Templates: Syntactic templates aligned with Ukrainian grammar were used to infer relations. For example: "повідомив про" maps to reports, "з використанням" to uses, and "експлуатує вразливість" to exploits.

The same three-tier confidence scoring was applied identically to the Ukrainian outputs, with high confidence assigned only when the LLM's semantic extraction agreed with at least one of the language-specific enhancement mechanisms.

4.1.3. Entity and Relation Normalization

The final stage involved rigorous post-processing to prepare the data for graph construction and alignment:

- Standardization: Entity text underwent standardized case and whitespace normalization;
- Deduplication: Fuzzy string matching from rapidfuzz library [29] was applied to merge near-duplicate or orthographically varied mentions of the same entity across articles and across languages;
- Predicate Normalization: The relational phrases extracted by the LLM were mapped to a canonical set of predicates. E.g., "reported," "reporting," and "повідомив про" were all normalized to the single edge type: reports;
- Consistent Identifiers: Every unique, normalized entity was assigned a global, language-prefixed ID, for example en_E1, uk_E1, which was important for the subsequent steps of cross-lingual clustering and alignment.

This dual-pipeline, high-confidence strategy ensured that the two generated knowledge graphs were not only highly accurate but also structurally and semantically consistent, allowing for reliable downstream analysis.

4.2. Knowledge Graph Construction

For algorithmic experimentation and semantic clustering, a local graph was built in Python using igraph [30]. Each node represented a normalized entity, and edges encoded extracted relations. This lightweight, in-memory graph allowed:

- Rapid prototyping and embedding-based computations;
- Easy integration with downstream clustering and NLP pipelines;
- Export of results to external formats for subsequent analysis.

For visual inspection and manual exploration, the previously built graph was utilized and the same entities and relations were exported them into a Neo4j graph database [31]. This made it possible to:

- Use Cypher queries to filter, merge, and analyze subgraphs;
- Visualize relational clusters and their interconnections interactively;
- Support bilingual inspection by keeping language-specific node labels and shared semantic identifiers.

This hybrid approach provided the flexibility of Python's analytical stack without redundancy in data processing, while igraph was used to build and analyze graph structures before exporting them for interactive exploration in Neo4j. The full set of normalized relation predicates used to describe cybersecurity events includes:

- reports (ORG reports VULNERABILITY)
- uses (ACTOR uses MALWARE)
- targets (MALWARE targets ORG)
- exploits (ACTOR exploits VULNERABILITY)
- affects (VULNERABILITY affects PRODUCT)
- part_of (INDICATOR part_of MALWARE)
- linked_to / related_to (for general association)

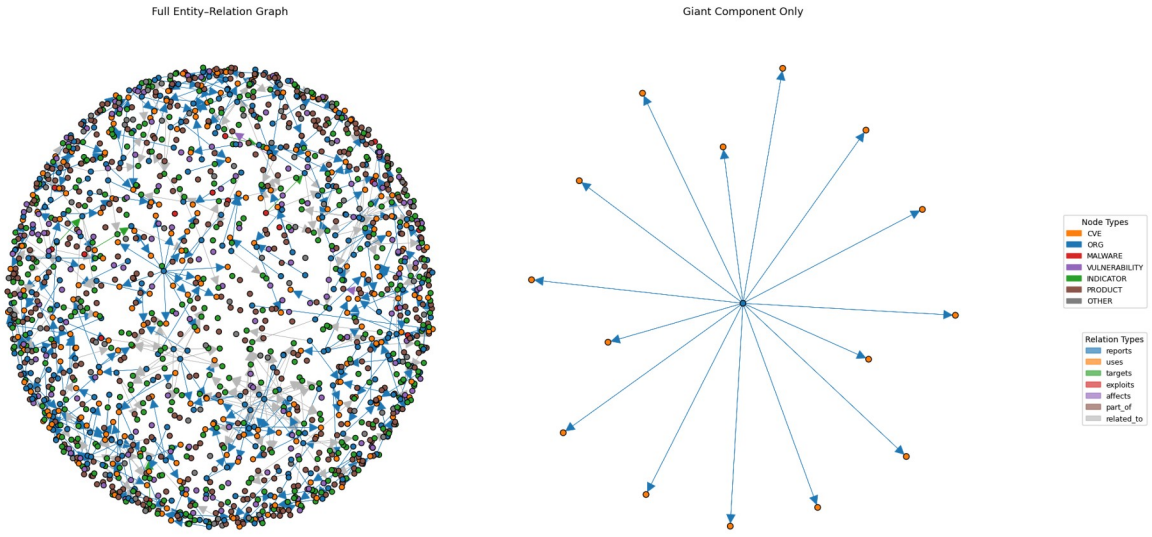

Figure 1: Knowledge graph created based on English Corpus: Full Graph and Giant Component Only.

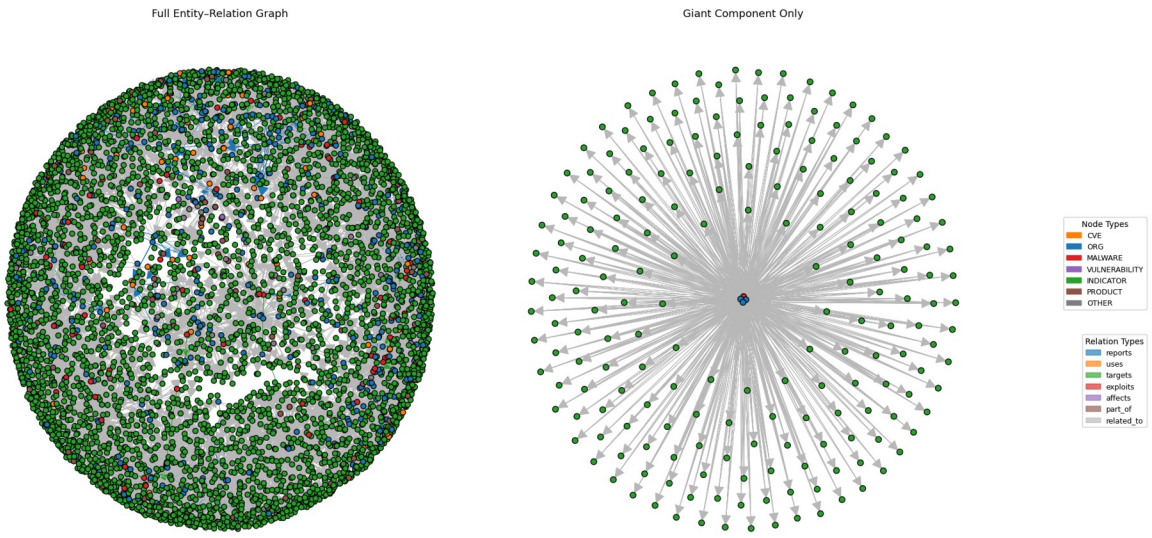

Figure 2: Knowledge graph created based on Ukrainian Corpus: Full Graph and Giant Component Only.

Data ingestion utilized Neo4j's LOAD CSV utility combined with Cypher queries to transform the extracted tabular data into the graph model. Entities were merged based on their names to ensure uniqueness, and relations were established by matching subject and object entities and creating the directed edges with the appropriate semantic predicate. To visualize the resulting graphs, two classical force-directed layout algorithms were employed: Fruchterman–Reingold [32] and Kamada–Kawai [33]. Both methods arrange nodes in two-dimensional space by simulating physical systems where edges act as springs that attract connected nodes, while unconnected nodes repel each other to minimize overlap and improve readability. The Fruchterman–Reingold algorithm emphasizes local clustering and community density, making it suitable for revealing tightly connected subgraphs, whereas Kamada–Kawai focuses on preserving global pairwise distances, producing layouts that more accurately reflect overall network topology. The use of both methods provided complementary perspectives on the structure of the cybersecurity knowledge graphs, highlighting local community formation and global connectivity patterns across entities and relations. This structure supports efficient querying and complex graph analytics. Interactive graph exploration and validation were performed using Neo4j Bloom. For both languages separated graph was created, as shown in Fig. 1 and Fig 2.

Fig. 1 presents the structure of a knowledge graph constructed from English corpus, where nodes represent extracted entities and edges denote semantic relations identified between them. The complete entity–relation graph, was arranged using the Kamada-Kawai layout to achieve an energy-minimized, interpretable distribution of nodes. The resulting topology reveals a highly fragmented structure composed of many small, weakly connected subgraphs and several denser clusters, reflecting the heterogeneity of information extracted from heterogeneous cybersecurity texts.

The right panel shows the giant component, defined as the largest connected subgraph in the network. This subgraph exhibits a hub-and-spoke configuration, in which a small number of central entities maintain most of the relational connections. Such organization suggests that the majority of entities are context-specific or appear in limited local narratives, while a few key entities serve as the primary relational anchors across corpora.

Fig 2. illustrates the structure of the graph built from Ukrainian corpus. The resulting structure exhibits a pronounced dominance of INDICATOR entities and relatively fewer cross-category connections, producing a highly fragmented topology with many small or weakly connected

components. This pattern reflects the narrower scope and less uniform relational coverage of entities mentioned in Ukrainian corpora compared to the English dataset.

The right panel highlights the giant component. It reveals a hub-centric organization, where a few core entities act as central connectors to a large number of indicators. This emphasizes the asymmetry in relational density, where many indicators are linked to a small number of central entities, consistent with reporting structures driven by incidents.

A comparison of the English and Ukrainian graphs reveals notable structural and semantic asymmetries between the two. The English graph demonstrates a more heterogeneous and interconnected topology, characterized by multiple medium-sized clusters containing diverse entity types. This structure suggests broader narrative coverage, with entities frequently linked through various relation types, reflecting the typically more comprehensive and standardized reporting style of English cybersecurity sources. In contrast, the Ukrainian graph exhibits a more centralized and fragmented organization. The distribution of node types is heavily skewed toward INDICATOR, while other entity categories appear sparsely connected. The giant component in the Ukrainian network follows a hub-and-spoke pattern, dominated by a few central entities connected to many peripheral indicators. This indicates that Ukrainian reports tend to emphasize specific incidents or campaigns, focusing on detailed technical observables rather than extensive inter-entity relationships. Consequently, the overall relational density and between types connectivity are lower compared to the English corpus, which may reflect differences in available data volume, linguistic expression, and practices of information sharing across reporting ecosystems.

Together, these findings underscore the linguistic and contextual variability in cybersecurity intelligence sources and highlight the importance of multilingual alignment and normalization when constructing unified knowledge graphs from heterogeneous report collections.

4.3. Semantic Entities Clustering

Each entity in the knowledge graph was embedded into a multilingual semantic space and grouped into dense clusters that capture recurring cyber-security themes.

To represent entity semantics, the LaBSE multilingual transformer model from the SentenceTransformers library, sentence-transformers/LaBSE, was used. This model produces high-quality, 768-dimensional sentence embeddings that are aligned across languages, which allowed to later perform cross-lingual alignment between English and Ukrainian clusters.

Each node's textual label was encoded into a vector with this model. For unsupervised grouping HDBSCAN [34], a density-based algorithm that automatically discovers clusters of variable size without needing to pre-specify the number of groups, was utilied.

The parameters were tuned to prioritize compact, semantically coherent clusters:

- minimal cluster size was set to 5;
- euclidean distance was utilized as metric;
- Automatic detection of outliers.

HDBSCAN was chosen over alternatives because it:

- can handle variable-density data;
- identifies both well-formed thematic groups and ambiguous/noisy nodes;
- does not require manual tuning of cluster count.

Each resulting cluster was assigned a unique semantic identifier and linked back to its original nodes in the NetworkX graph. Cluster naming was performed automatically using a heuristic token-frequency approach. For each cluster, entity labels were aggregated and the top three most frequent words were extracted to generate concise, human-readable names, for example:

"Apple_Frameworks_Vulnerabilities", *"Malware_Distribution_Tools"*, *"APT_Groups"*. These names served as preliminary descriptors for downstream model-based labeling.

To quantitatively assess clustering quality, several internal metrics were computed for each entity type, as shown in Table 1. To interpret cluster topology, the high-dimensional embeddings were reduced to 2D using UMAP with parameters n_neighbors=15, min_dist=0.1, metric='cosine'. The resulting maps were visualized, where each point corresponds to an entity and color denotes the assigned cluster. This allowed visual confirmation of dense semantic regions representing coherent cybersecurity domains such as:

- Vulnerability families;
- Malware or exploit toolsets;
- Organizations and actors;
- Indicators of compromise.

Table 1
Metrics, used for clustering evaluation

Metric	Description
Silhouette Score	Measures the separation between clusters (−1 to 1). Higher is better.
Mean Intra-Cluster Cosine Similarity	Average semantic similarity within clusters.
Noise Ratio	Fraction of unclustered (outlier) entities.
Average Cluster Size	Average number of entities per cluster.

4.4. Model-based Cluster Labeling and Alignment

After clustering, three large language models were employed to generate semantic labels for each cluster:

- Meta-LLaMA-3-8B-Instruct [35];
- Mistral-7B-Instruct-v0.3 [36];
- Falcon-7B-Instruct [37].

Each model received a short prompt with a list of representative entities from a cluster and was asked to produce a 3-6-word human-readable label describing the shared theme. Each model's output was compared across clusters using Jaccard similarity [38] and manual review to assess agreement. If all models converged on similar semantics, the cluster was marked as high-confidence. Discrepancies were analyzed to identify systematic differences in model interpretation. For each cluster, model labels were merged into a meta-label, representing consensus across LLMs. A reliability weight was assigned based on:

- lexical similarity between model outputs;
- coherence of labels with the original entity context;
- consistency of reasoning patterns in the model output.

4.5. Cross-Lingual Alinment

After generating high-confidence clusters and meta-labels for both English and Ukrainian corpora, the next step was to align them into a unified bilingual structure. Many cybersecurity reports are released in parallel by Ukrainian and international agencies. Although they often refer to the same events or threat actors, their entity names and descriptions may differ linguistically. To consolidate this knowledge, semantically equivalent clusters across both languages were aligned into shared meta-clusters.

Each English cluster C_{en} and Ukrainian cluster C_{uk} was represented by a feature vector combining:

- Semantic Centroid Similarity: mean LaBSE embedding of all entities within the cluster. Pairwise cosine similarity between centroids provided a semantic alignment score;
- Translated meta-label Overlap: meta-labels were automatically translated, and lexical Jaccard similarity was computed;
- Shared Entities: Some entity identifiers are language-invariant and were used as hard anchors between clusters.

The combined similarity score S for each pair was computed as:

$$S = \alpha \cdot \text{cosine}(E_{en}, E_{uk}) + \beta \cdot \text{Jaccard}(L_{en}, L_{uk}) + \gamma \cdot \text{EntityOverlap} \quad (1)$$

where $\alpha=0.6$, $\beta=0.3$, and $\gamma=0.1$ provided balanced weighting between semantic and lexical similarity.

Cluster pairs with $S > 0.75$ were treated as bilingual equivalents, while scores between $0.5 < S \leq 0.75$ indicated possible relatedness requiring manual validation. Lower-scoring pairs were retained in language-specific subsets.

Traditional multilingual knowledge graph alignment methods, such as MTransE [39] and its variants, primarily rely on structural embeddings and pre-existing interlingual links to project entities from different languages into a shared space. While effective for well-aligned graphs with rich cross-lingual mappings, such models struggle with semantically heterogeneous or partially overlapping corpora like cybersecurity reports. In contrast, approach proposed in this work employs LaBSE-based semantic embeddings combined with label similarity to achieve thematic rather than purely structural alignment. This enables the discovery of conceptually equivalent clusters even in the absence of direct correspondences on entitylevel, extending alignment beyond the limitations of models like MTransE.

4.6. Meta-Cluster Analysis and Association Rules

To uncover broader patterns across the unified dataset, association rule mining was applied to the merged meta-cluster dataset. Each meta-cluster was treated as a transaction containing its associated entities. Using the Apriori algorithm [40] from mlxtend.frequent_patterns [41], frequent co-occurrence patterns between entities and cluster themes were identified. Rules were generated based on minimum support (0.1) and lift (>1.0) thresholds. Example of rule generation is presented in Table 2.

Table 2
Association Rule generation example

Rule	Support	Confidence	Lift
{phishing, credential theft} : {APT28}	0.21	0.78	1.42

	0.17	0.66	1.3
{CVE-2023-38831, WinRAR} : {exfiltration}			

These rules highlight consistent relationships between malware families, exploited CVEs, and operational tactics across reports in both languages. Moreover, since association rules naturally form directed relationships between entities, they can be incorporated back into the knowledge graph as inferred edges, extending the framework toward dynamic, self-enriching graph-based reasoning.

5. Results and Discussion

5.1. Clustering Results

Clustering was evaluated in two ways: visually and quantitive. Fig. 3 and Fig. 4 present the visual results of clustering in both languages. Fig. 3. shows, that entity types with a clear conceptual hierarchy, such as ORG and PRODUCT form tighter, fewer clusters. At the same time, highly technical or unique entity types like CVE and INDICATOR result in a high number of fragmented, isolated, or unclustered groups, reflecting their lack of generalized semantic context.

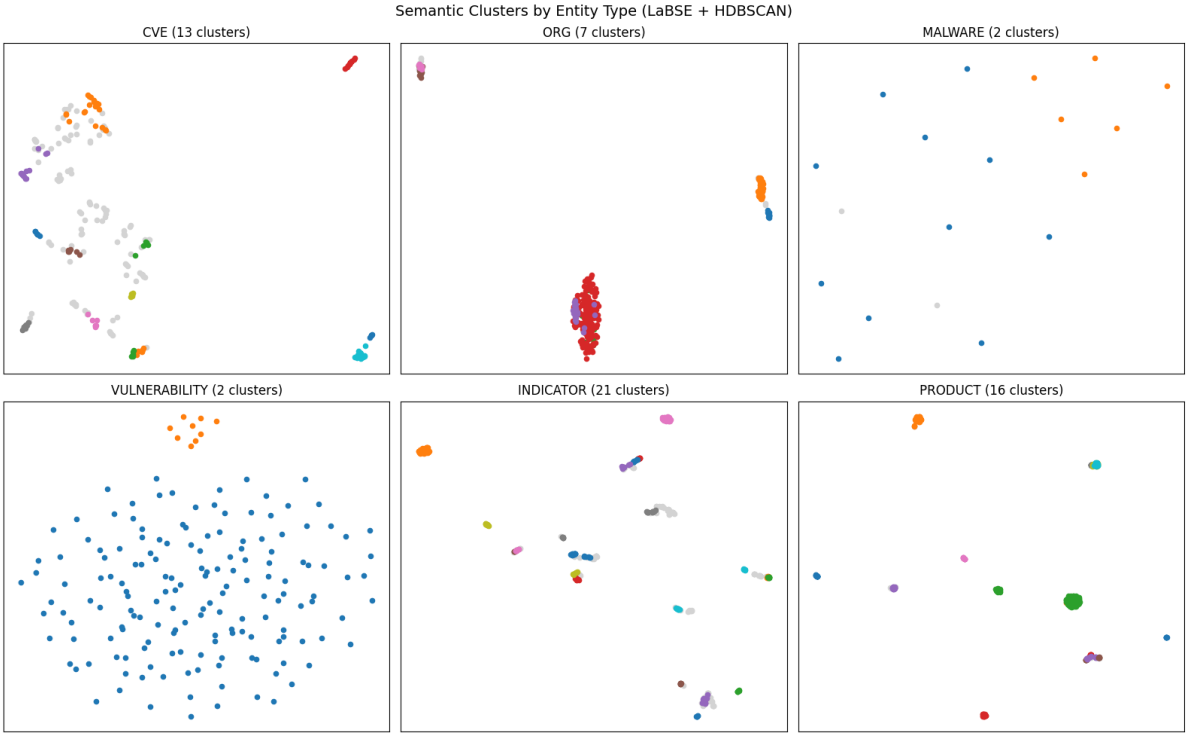

Figure 3. English corpus clustering results.

Figure 4. Ukrainian corpus clustering results.

The clustering topology of the Ukrainian corpus shown in Fig. 4. reflects a threat landscape dominated by local and highly targeted activity. The large core in INDICATOR entities and the low cluster count in CVEs suggest that reporting often concentrates on a narrow set of high-priority events. The high cluster count in ORG entities shows that the most semantically unique concepts are often tied to specific local institutions, demonstrating a focus on the geopolitical context of the reporting. The complete absence of clusters for VULNERABILITY entities suggests that the generic concepts are often mentioned without sufficient surrounding unique descriptive text to create clear thematic groups, possibly due to a focus on the immediate IoC or the specific CVE ID instead.

The application of LaBSE embeddings and HDBSCAN clustering to both the English and Ukrainian cybersecurity corpora revealed major differences in how key entities are structured and reported across the two linguistic and geopolitical contexts.

The ORG entity type exhibits a direct inversion in separation between the two corpora. The English corpus yielded a low number of large clusters, suggesting that international reporting frequently mentions a few major global organizations which share a common semantic context. In contrast, the Ukrainian corpus showed significant fragmentation. This high count indicates the model effectively isolates many smaller, highly specific groups representing domestic state entities and critical infrastructure targets, which are semantically distinct from the broader global context, reflecting a highly localized threat intelligence priority.

Table 3
Quantitative clustering evaluation

Entity type	Nodes		Clusters		Silhouette		Intra-sim		Noise points		Noise ratio	
	EN	UA	EN	UA	EN	UA	EN	UA	EN	UA	EN	UA
CVE	229	74	13	4	0.534	0.372	0.952	0.898	113	26	0.493	0.351
ORG	254	438	7	14	0.968	0.932	0.982	0.968	10	48	0.039	0.11

MALWARE	19	93	2	6	0.510	0.802	0.820	0.929	2	29	0.105	0.312
VULNERABILITY	169	9	2	0	1.000	-	1.000	-	0	9	0.0	1.0
INDICATOR	398	4254	21	13	0.740	0.202	0.906	0.781	126	238	0.317	0.056
PRODUCT	469	107	16	6	0.944	0.938	0.974	0.946	13	3	0.028	0.028

The categories related to system flaws such as CVE and VULNERABILITY show a difference in reporting depth. The English corpus produced a high number of CVE clusters, reflecting global reporting on a wide variety of specific vendor patches and exploitation vectors. Conversely, the Ukrainian corpus collapsed into only four CVE clusters, suggesting reports are focused on a narrower set of actively exploited vulnerabilities relevant to regional campaigns. Furthermore, the complete absence of any VULNERABILITY clusters in the Ukrainian corpus, compared to two in English, suggests a lack of rich, descriptive semantic context around generic vulnerability terms, where the reporting priority often skips directly to the specific CVE identifier or the immediate IoC.

The INDICATOR entities reveal how the reporting methodology affects semantic density. The English corpus produced a large number of fragmented clusters, which is expected for globally diverse, unique artifacts like hashes and IP addresses. The Ukrainian corpus, however, consolidated most of its indicators into one massive, dense core cluster. This indicates that the majority of IoCs in Ukrainian reporting are highly co-occurring and likely derived from similar, repetitive technical alerts, leading to a high semantic density around a limited number of recurring artifact sets.

The clustering of MALWARE entities also points to a difference in focus. The English corpus favored a lower number of broad clusters, suggesting a high-level semantic distinction. The Ukrainian corpus, however, showed a more fragmented structure with six clusters. This suggests the regional threat landscape involves tracking a greater variety of distinct, locally relevant, or less known malware families that do not fit neatly into the two global categories.

In summary, the English KG structure is characterized by high fragmentation of technical details and high cohesion of organizational context. The Ukrainian KG structure is characterized by high cohesion of technical details and high fragmentation of organizational context, underscoring the shift in semantic priority from a broad global landscape to a focused, localized, and technically repetitive regional conflict.

Table 3 shows the results of quantitative evaluation of clustering. The English corpus produced more compact and separated clusters across all entity types, reflected by higher silhouette and intra-cluster similarity values. The Ukrainian corpus, while smaller and more heterogeneous, still achieved high coherence in ORG, MALWARE, and PRODUCT categories, confirming the stability of the multilingual embeddings.

The drop in INDICATOR silhouette for Ukrainian data can be attributed to the overwhelming number of short, highly contextual indicators which exhibit weak semantic structure in embedding space. Conversely, VULNERABILITY entities clustered perfectly in English, indicating strong internal similarity and clear separability.

English clusters exhibited generally low noise ratios, particularly for ORG and PRODUCT, implying high semantic consistency and frequent co-occurrence across documents. In contrast, categories such as CVE and MALWARE displayed elevated noise values, corresponding to their diverse naming conventions and rapidly evolving terminology. The Ukrainian dataset showed similar trends but with slightly higher overall noise, which can be attributed to smaller corpus size and more heterogeneous phrasing. These noise indicators thus serve as diagnostic signals, highlighting areas where additional data or improved normalization could further enhance clustering robustness.

5.2. Meta-Label Generation & Adjustement

After the initial clustering stage, three LLMs independently proposed short descriptive labels for each English cluster. Although most labels captured correct topical semantics, lexical inconsistency and overlapping formulations limited their direct use for bilingual alignment. To address this, a human-in-the-loop revision step was applied, focusing on normalizing terminology, resolving ambiguity, and harmonizing phrase length and style.

The effect of human refinement was measured by recalculating bilingual alignment scores using automatic versus adjusted by human meta-labels and it is shown in Table 4.

Table 4
Effect of adjusting the clusters meta-labels.

Adjustment Type	Mean Similarity	High Confidence	Max Similarity
Automatic	0.47	2/13 (15%)	0.685
Human-Adjusted	0.53	3/13 (23%)	0.786

This demonstrates that even minimal human editing of LLM outputs improves cross-lingual cluster alignment by approximately 13 % in mean similarity and increases the number of confident bilingual matches. Human adjustment primarily improved cases where LLMs used overly narrow or specific phrasing. In contrast, clusters labeled with clear vendor or vulnerability identifiers already aligned well automatically. This confirms that human and machine hybrid labeling achieves higher semantic stability across languages while preserving efficiency.

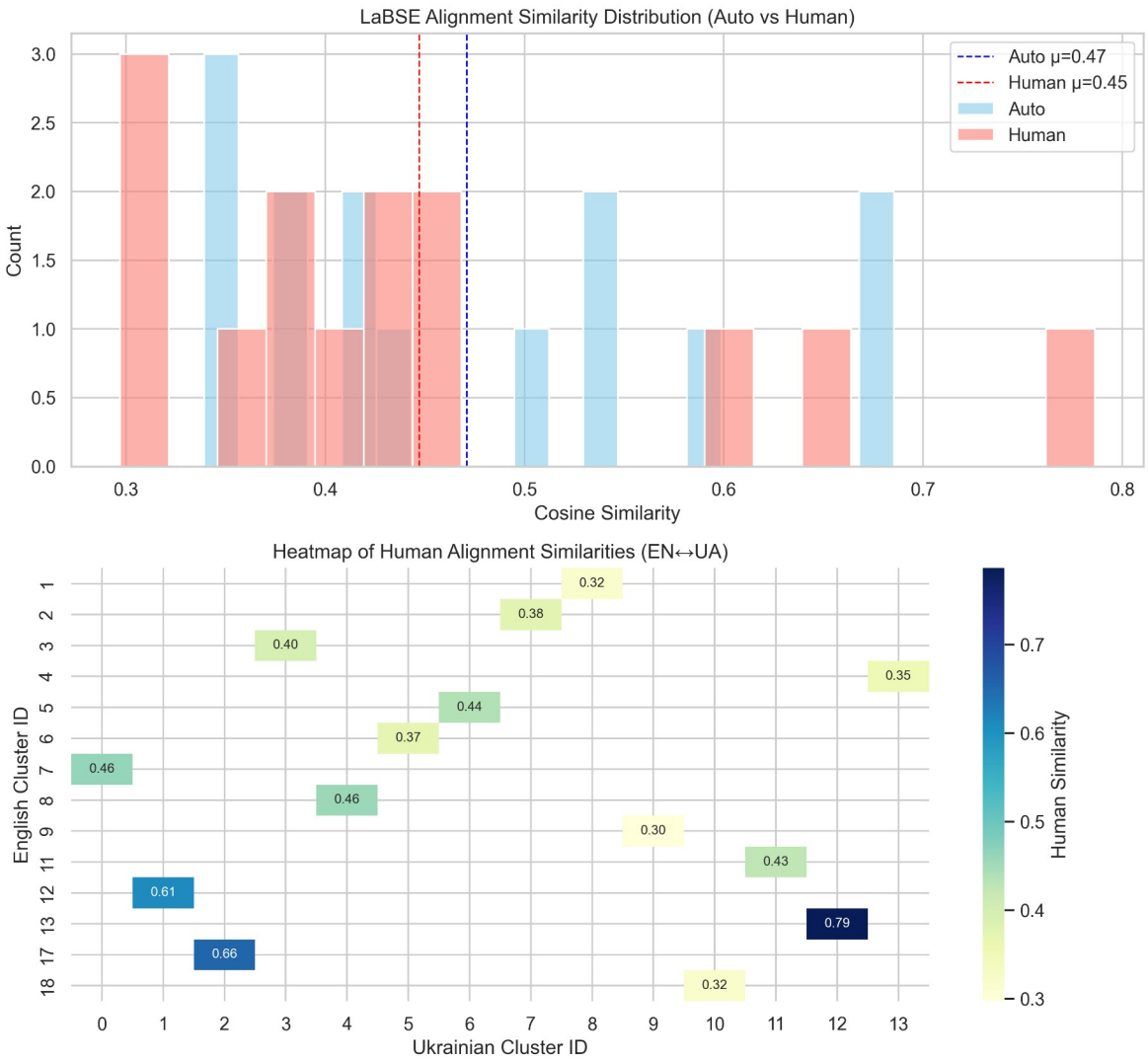

Figure 5. Comparison of Auto vs. Human Alignment

Figure 5 illustrates the effect of manual meta-label refinement on cross-lingual alignment quality. The upper panel compares the distribution of cosine similarity scores between automatically generated and adjusted by human meta-labels. The dashed vertical lines indicate the mean similarity for each approach: auto = 0.47, human = 0.45. While both follow similar distributions in the 0.3-0.5 range, human-adjusted alignments show a right-tail shift above 0.6, indicating more high-confidence bilingual matches.

The lower panel visualizes the pairwise similarity heatmap between English and Ukrainian clusters after human correction. Darker cells correspond to stronger semantic correspondence, with peak alignment observed between clusters describing network configuration scripts and network infrastructure settings. Together, these results confirm that normalization by human improves the consistency of cluster semantics and increases the number of bilingual matches, particularly for technical domains where lexical differences between languages are greatest.

5.3. Bilingual Graph Integration

To evaluate the effectiveness of the cross-lingual alignment pipeline, both language-specific knowledge graphs were merged into a unified graph. Each node retained its original language tag, while edges representing semantically aligned clusters were added based on the computed LaBSE

similarity scores. Fig. 6 shows the resulting bilingual knowledge graph visualized in Neo4j. In this visualization:

- Nodes are displayed in a uniform color, highlighting the semantic unification across languages.
- Blue edges represent original intra-language relations, extracted independently from the TALOS and CERT-UA corpora.
- Orange edges represent cross-lingual links generated during the alignment phase, connecting semantically equivalent or contextually related entities.

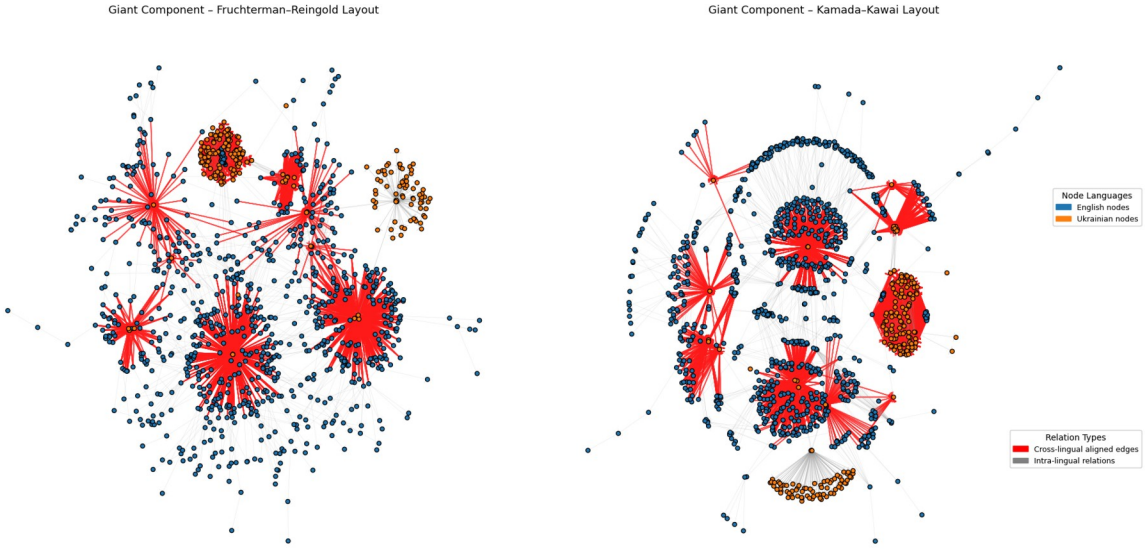

Figure 6. Bilingual Cybersecurity Knowledge Graph

Fig. 6 presents the bilingual alignment of the English and Ukrainian graphs, showing the giant component after alignment. Each node represents an entity extracted from cybersecurity reports, with blue nodes denoting English entities and orange nodes representing Ukrainian entities. Edges are of two kinds: gray lines indicate the original relations within each language, while red lines represent the cross-lingual alignment links that connect semantically equivalent or related entities across the two corpora.

The left panel employs the Fruchterman-Reingold layout, which emphasizes local clustering and community density, while the right panel applies the Kamada-Kawai layout, which better preserves global structural relationships. Both layouts reveal multiple bilingual clusters where entities from both languages are tightly coupled by alignment edges, forming distinct hubs of shared cybersecurity knowledge. At the same time, the persistence of several monolingual regions with few or no red edges indicates incomplete alignment or concepts specific in language or unique to one reporting domain.

Overall, the visualization highlights how cross-lingual linking unites previously separate semantic spaces, creating a connected bilingual knowledge structure while also exposing gaps and asymmetries in entity coverage between the English and Ukrainian cybersecurity narratives.

The aligned bilingual graph provides information about how English and Ukrainian cybersecurity reporting intersect and diverge at the entity level. Whereas the monolingual graphs exhibited distinct internal characteristics, the aligned graph integrates these structures through explicit cross-lingual correspondence edges.

The dense red connections observed in several regions represent successful semantic alignment of equivalent entities, such as widely reported vulnerabilities, malware strains, or organizations active across both corpora. These alignments demonstrate overlap in core threat intelligence

concepts, confirming that both languages describe many of the same cybersecurity phenomena. However, the uneven distribution of red links and the presence of isolated monolingual clusters suggest persistent linguistic and contextual fragmentation. Certain Ukrainian entities have no English counterpart, likely reflecting locally focused incident reporting, unique terminology, or the use of transliterated names not captured during entity normalization. Conversely, some English clusters lack Ukrainian connections, indicating information asymmetry and differing dissemination patterns in global versus regional cybersecurity reporting.

Altogether, the aligned structure reveals that cross-lingual entity mapping not only improves the connectivity and completeness of multilingual cybersecurity graphs but also provides a powerful lens for analyzing semantic coverage gaps and regional biases in threat intelligence data.

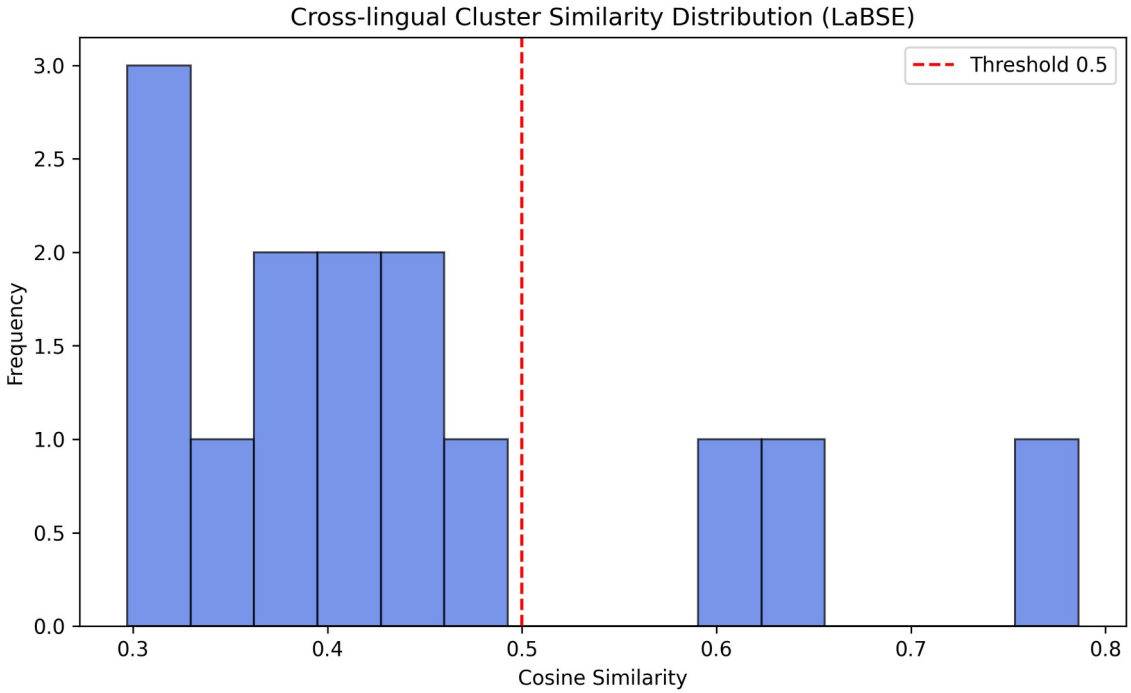

Figure 7. Cross-lingual Cluster similarity Distribution.

The similarity distribution shown in Fig. 7 provides a quantitative perspective on the strength of the bilingual alignment achieved by the LaBSE model. Most English-Ukrainian cluster pairs exhibit cosine similarity between 0.3 and 0.5, indicating moderate cross-lingual relatedness where topic domains partially overlap but surface terminology differs. A smaller but meaningful subset of cluster pairs surpasses the 0.5 threshold, signifying strong semantic correspondence; these typically involve shared identifiers such as CVE codes, malware family names, or vendor organizations that appear consistently across both corpora. The right-hand tail of the distribution captures the alignments with the highest confidence, confirming that the embedding-based approach can successfully recognize conceptually equivalent topics even in linguistically divergent texts. The relatively low density below 0.3 suggests that false alignments were effectively filtered out, emphasizing the discriminative power of the threshold. Overall, the histogram supports the visual evidence from Fig. 6 showing that the bilingual integration preserved graph coherence while introducing meaningful cross-language bridges concentrated in semantically stable cybersecurity domains.

5.4. Association Rule Analysis

To uncover recurrent relationships among cybersecurity entities and thematic clusters, association rule mining was applied to the results of the semantic clustering and meta-label alignment stages. Each cluster was treated as a transaction containing its constituent entities. The Apriori algorithm

from the mlxtend.frequent_patterns module was used to identify frequent co-occurrence patterns across these transactions.

Rules were generated under the following parameters:

- Minimum Support: 0.1 - ensuring that patterns appear in at least 10% of the clusters;
- Minimum Lift: 1.0 - retaining only those associations that occur more frequently than chance;
- Metric: “lift” for rule evaluation, with confidence reported as an auxiliary indicator.

This process was executed independently for each dataset and for the bilingual aligned meta-clusters, allowing comparative assessment of cross-lingual relational coherence.

The English dataset, while highly structured and semantically coherent, produced no association rules under the specified thresholds. This absence of frequent itemsets indicates high thematic purity and low inter-cluster overlap, consistent with the earlier clustering metrics showing high silhouette coefficients and strong intra-cluster similarity. In practice, English reports focus on distinct, non-overlapping vulnerability domains, which limits the statistical co-occurrence of entities across clusters.

Conversely, the Ukrainian dataset exhibited several rules with high support, reflecting richer cross-cluster linkages. The overlap between indicators, vulnerabilities, and organizations indicates denser semantic interconnection typical for situational threat reporting. These associations highlight the operational relationships between malware campaigns, exploited vulnerabilities, and adversarial tactics, revealing the characteristic co-occurrence of named entities within Ukrainian threat reports.

After the bilingual alignment step, association rule mining uncovered interlingual relationships, demonstrating how combining corpora in two languages enriches the resulting knowledge representation. The merged meta-cluster dataset produced higher rule density and support, confirming that cross-lingual integration compensates for monolingual sparsity. Frequent rules connected semantically equivalent entities across languages, validating the effectiveness of the alignment pipeline.

Together, these findings show that while English sources provide clear, isolated threat descriptions, Ukrainian and cross-lingual corpora enable contextual synthesis and higher-order relational inference, which are essential for multilingual threat intelligence analysis.

6. Conclusions

This study introduced and validated a framework for constructing and analyzing bilingual cybersecurity knowledge graphs that integrate large language models, multilingual embeddings, and analytics based on graphs. The framework effectively extracts entities and relations from English and Ukrainian corpora, organizes them into semantically coherent clusters, and aligns conceptually equivalent structures across languages. Quantitative and visual evaluations confirm that the proposed approach produces compact, interpretable clusters and meaningful cross-lingual alignments, revealing both shared and patterns specific for language in cybersecurity reporting.

The application of LaBSE embeddings and HDBSCAN clustering ensured robust cross-lingual compatibility and unsupervised discovery of latent topics, while LLM-based labeling enhanced semantic interpretability. The integration of English and Ukrainian graphs through bilingual alignment enabled the identification of conceptually equivalent entities and enriched the overall contextual structure of the combined knowledge graph. Association rule mining further demonstrated the ability of the framework to uncover higher-order semantic relationships, providing insight into recurrent threat patterns across languages. These rules can also be represented as directed edges within a secondary graph layer, extending the knowledge graph with inferred associations.

Beyond its methodological contributions, this research highlights the potential of multilingual graph-based analytics for improving cyber threat intelligence. Future work will extend the corpus size, incorporate additional languages, and explore in real time ingestion of threat reports to support dynamic, multilingual monitoring of emerging cyber threats.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to: translate certain text fragments into English, perform grammar and spelling checks, and paraphrase or reword content. After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Barnum. "Standardizing cyber threat intelligence information with the structured threat information expression (stix)." Mitre Corporation 11 (2012): 1-22.

[2] J. Connolly, M. Davidson, C. Schmidt. "The Trusted Automated eXchange of Indicator Information (TAXII)." The MITRE Corporation (2014): 1-20.

[3] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, C. B. Thomas. "MITRE ATT&CK: Design and Philosophy." The MITRE Corporation (2018).

[4] B. M. Pavlyshenko. "AI Approaches to Qualitative and Quantitative News Analytics on NATO Unity." arXiv preprint (2025). https://doi.org/10.48550/arXiv.2505.06313

[5] B. M. Pavlyshenko. "Multilevel Analysis of Cryptocurrency News using RAG Approach with Fine-Tuned Mistral Large Language Model." arXiv preprint (2025). https://doi.org/10.48550/arXiv.2509.03527

[6] B. M. Pavlyshenko. "Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model." arXiv preprint (2023). https://doi.org/10.48550/arXiv.2309.04704

[7] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, A. Jain. "Structured information extraction from scientific text with large language models." Nature Communications 15(1) (2024): 1418. https://doi.org/10.1038/s41467-024-45563-x

[8] A. Goel, A. Gueta, O. Gilon, C. Liu, S. Erell, L. H. Nguyen, X. Hao, B. Jaber, S. Reddy, R. Kartha, J. Steiner, I. Laish, A. Feder. "LLMs Accelerate Annotation for Medical Information Extraction." Proceedings of the 3rd Machine Learning for Health Symposium (2023): 82-100.

[9] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, H. Wang. "Large Language Models for Cyber Security: A Systematic Literature Review." ACM Trans. Softw. Eng. Methodol. (2025). https://doi.org/10.1145/3769676

[10] J. Zhang, H. Bu, H. Wen, Y. Liu, H. Fei, R. Xi, L. Li, Y. Yang, H. Zhu, D. Meng. "When LLMs meet cybersecurity: A systematic literature review." Cybersecurity 8(1), 55 (2025). https://doi.org/10.1186/s42400-025-00361-w

[11] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, H. Wang. "Large Language Models for Cyber Security: A Systematic Literature Review." ACM Transactions on Software Engineering and Methodology (2025). https://doi.org/10.1145/3769676

[12] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al. "Language Models are Few-Shot Learners." Advances in Neural Information Processing Systems (NeurIPS) 33 (2020): 1877-1901. Available at: https://arxiv.org/abs/2005.14165

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805 (2018). https://doi.org/10.48550/arXiv.1810.04805

[14] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang. "Language-agnostic BERT Sentence Embedding." arXiv preprint arXiv:2007.01852 (2020). https://doi.org/10.48550/arXiv.2007.01852

[15] M. A. Shah, M. J. Iqbal, N. Noreen, I. Ahmed. "An automated text document classification framework using BERT." Int. J. Adv. Comput. Sci. Appl. 14(3) (2023): 279-285.

[16] R. Savant, A. Shelke, S. Todmal, S. Kanphade, A. Joshi, R. Josh. "Universal cross-lingual text classification." In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (2024): 1-6.
[17] A. Adhikari, A. Ram, R. Tang, J. Lin. "Docbert: Bert for document classification." arXiv preprint arXiv:1904.08398 (2019). https://doi.org/10.48550/arXiv.1904.08398
[18] F. A. Galatolo, G. Martino, M. G. Cimino, C. O. Tommasi. "Dense Information Retrieval on a Latin Digital Library via LaBSE and LatinBERT Embeddings." In: DATA (2023): 518-523.
[19] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang. "Language-agnostic BERT sentence embedding." arXiv preprint arXiv:2007.01852 (2020). https://doi.org/10.48550/arXiv.2007.01852
[20] H. Al-Jabri, A. Al-Harrasi, M. Al-Jabri. "From Text to Actionable Intelligence: Automating STIX Entity and Relationship Extraction." arXiv preprint arXiv:2507.16576 (2025). https://doi.org/10.48550/arXiv.2507.16576
[21] T. Muennighoff, K. Enevoldsen, J. Muennighoff. "Examining Multilingual Embedding Models Cross-Lingually Through LLM-Generated Adversarial Examples." arXiv preprint arXiv:2502.08638 (2025). https://doi.org/10.48550/arXiv.2502.08638
[22] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, A. Jain. "Structured information extraction from scientific text with large language models." Nature Communications 15(1) (2024): 1418. https://doi.org/10.1038/s41467-024-45563-x
[23] W. Chen, H. Li, Z. Yang, R. Zhang. "Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment." In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 1478–1484 (2017).
[24] CERT-UA. "Official Website of the Computer Emergency Response Team of Ukraine." Available at: https://cert.gov.ua/ (Accessed: October 12, 2025).
[25] Cisco Talos. "Talos Intelligence Blog." Available at: https://blog.talosintelligence.com/ (Accessed: October 12, 2025).
[26] Microsoft. "Playwright: Fast and reliable end-to-end testing for modern web apps." Available at: https://playwright.dev/ (Accessed: October 12, 2025).
[27] L. Richardson. "Beautiful Soup: Python library for pulling data out of HTML and XML files." Available at: https://www.crummy.com/software/BeautifulSoup/ (Accessed: October 12, 2025).
[28] OpenAI. "GPT-5 System Card." Available at: https://cdn.openai.com/gpt-5-system-card.pdf (Accessed: October 12, 2025).
[29] Maxbachmann. "rapidfuzz: Fuzzing library that provides high performance fuzzy string matching." Available at: https://github.com/maxbachmann/rapidfuzz (Accessed: October 12, 2025).
[30] The igraph Core Team. The igraph Software Package. Available at: https://igraph.org/ (Accessed: October 13, 2025).
[31] Neo4j, Inc. Neo4j Graph Database. Available at: https://neo4j.com/ (Accessed: October 13, 2025).
[32] T. M. J. Fruchterman and E. M. Reingold. "Graph drawing by force-directed placement." Software: Practice and Experience 21(11) (1991). https://doi.org/10.1002/spe.4380211102
[33] T. Kamada and S. Kawai. "An algorithm for drawing general undirected graphs." *Information Processing Letters* **31**(1) (1989): 7–15. https://doi.org/10.1016/0020-0190(89)90102-6
[34] L. McInnes, J. Healy, S. Astels. "hdbscan: Hierarchical density based clustering." Journal of Open Source Software 2(11), 205 (2017). https://doi.org/10.21105/joss.00205
[35] LLaMA Team, A. K. F., et al. "Llama 3: Open Foundation Models for a New Era." arXiv preprint (2024). https://doi.org/10.48550/arXiv.2407.21783
[36] A. Jiang, A. W. C., S. M. T., et al. "Mistral 7B." arXiv preprint (2023). https://doi.org/10.48550/arXiv.2310.06825
[37] G. J. J. T., E. P. B. A. M. A., et al. "The RefinedWeb Dataset for Falcon LLM: Permitting Further Open Pretraining." arXiv preprint (2023). https://doi.org/10.48550/arXiv.2311.16867
[38] P. Jaccard. "Étude comparative de la distribution florale dans une portion des Alpes et du Jura." Bulletin de la Société Vaudoise des Sciences Naturelles 37(140) (1901): 547-579.

[39] M. Chen, Y. Tian, M. Yang, C. Zaniolo. "Multilingual knowledge graph embeddings for cross-lingual knowledge alignment." arXiv preprint (2016). https://doi.org/10.48550/arXiv.1611.03954
[40] R. Agrawal, R. Srikant. "Fast algorithms for mining association rules." In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), pp. 487–499 (1994).
[41] S. Raschka. "MLxtend: Providing machine learning functionality for Python." Journal of Open Source Software 3(24), 631 (2018). https://doi.org/10.21105/joss.00631