Journal of Intelligent Information and Communication

Research Paper

Multilingual Security Document Understanding Based on XLNet Transfer Learning

Lesław Jura[1, *], Tadeusz Kacz[1] and Bogdan Kalisz[2]

[1] Faculty of Informatics, University of Gdansk, Gdansk, 80-952, Poland
[2] Faculty of Information Technology, University of Rzeszow, Rzeszow, 35-959, Poland
*Corresponding author: leslaw.jur@ug.edu.pl

Abstract. In order to address the technological issue of automatic multilingual security document interpretation, this study proposes a specific architecture based on XLNet-based transfer learning. First, we discovered that the real security materials contain a variety of languages as well as different degrees of structural irregularity and semantic complexity after conducting a thorough technical examination of them. Permutations, adaptive tokenization, domain-specific feature learning, and other techniques are the foundation of the suggested method of handling contexts. The model achieved a macro-averaged accuracy of 92.2% for English, 90.6% for Chinese, and maintained an accuracy of over 87% across all low-resource languages using a relatively large-scale, high-quality benchmark of over 120,000 annotated security papers in six languages. This structure has demonstrated lower entity boundary errors and higher F1 scores for rare and code-mixed event categories when compared to the well-known models of BERT and RoBERTa. It has been discovered that the model is rather stable in identifying threats and resolving ambiguity among the compliance and vulnerability descriptions based on the aforementioned thorough error analysis and real case validation. A new engineering standard for cross-lingual cybersecurity intelligence and compliance analysis has been established based on the aforementioned results, which show that permutation-driven transfer learning can accomplish dependable, high-precision multilingual information extraction and categorization.

Keywords: *Multilingual Document Processing, Security Text Analysis, Transfer Learning, XLNet*

Received on 11 December 2024, Accepted on 25 March 2025, Published on 03 April 2025

Introduction

Security documentation in various languages has increased by several orders of magnitude as a result of the widespread adoption of digital transformation in the government, finance, health, and energy sectors [1]. More thorough studies of incident reports, regulatory warnings, vulnerability notifications, and audit statements in many languages are currently being carried out due to the recent shift in cybersecurity and cross-border compliance norms [2]. However, both human specialists and conventional AI systems find it challenging to interpret these papers' intricate structure, technical jargon, and context-sensitive semantics [3]. There are several discrepancies between high-resource and low-resource languages, and information extraction and semantic understanding are also hampered by domain-specific words and policy variations [4]. As the number and variety of security documents continue to grow, it is more probable that certain employees will not fully review them; as a result, responses may be delayed or inaccurate, and regulatory requirements may not be fulfilled [5]. Traditional text mining and other rule-based, static machine learning techniques exhibit low scalability and adaptation to the numerous types of foreign security documents and linguistic factors [6]. In cross-lingual environments, recall, precision, and F1-score frequently decline dramatically; as a result, the existing technology is unable to satisfy practical needs [7]. Because of this, there has been a steady increase in demand for automated, reliable, and language-independent solutions to the security document intelligence challenge [8]. Both sophisticated language models and flexible transfer-learning backbones must be used to learn intricate cross-lingual and domain-specific features with little supervision in order to construct such

solutions [9]. These days, a lot of organizations in the governance and cybersecurity fields must also comply with these requirements [10].

Given the aforementioned issues, transfer learning has transformed natural language processing (NLP) by encouraging cross-domain and cross-lingual generalization, particularly for deep pre-trained language models [11]. For document-level representation, the permutation-based autoregressive model XLNet has shown good performance in handling context and flexible dependency modeling [12]. XLNet is a bidirectional encoder that can better utilize context at various points and does not have the issues associated with masked language modeling [13]. According to recent research, XLNet can effectively extract information and classify documents from structurally complicated, multilingual data, something that standard models frequently struggle with [14]. Despite the aforementioned noteworthy accomplishments, there are comparatively few systematic research on the application of XLNet-driven transfer learning to real-world multilingual security documents in low-resource languages, regulatory compliance, security threat identification, etc. [15].

This study proposes a comprehensive XLNet-based system for the understanding and processing of multilingual security documents in diverse linguistic and application settings in order to address this shortcoming. Robust experiments on real and synthetic datasets in representative security domains have been carried out, and an end-to-end method integrates strong tokenization, cross-lingual embeddings, domain adaptation, and task-specific fine-tuning. The following are the principal accomplishments: A flexible architecture for multilingual security document intelligence is introduced, together with thorough benchmarking against conventional baselines and fresh insights on mistake characteristics and useful applications. The remainder of this work will be structured to walk readers through the system design, experiments, analysis, and underlying research before providing an overview of the findings and future research directions.

Related Work

Multilingual Security Document Understanding

Growing international organizations increasingly need to understand and automatically analyze multilingual security papers [16]. The earliest types of translation, rule-based translation and manual annotation, have not yet resolved issues like the vast range of technical words and semantics in security documents [17]. People and conventional algorithms for security data analysis cannot consistently extract reports, guidelines, and advisories due to their varied styles [18]. Despite recent advancements in neural machine translation and cross-lingual word embeddings to address the issue of semantic disparities, their actual performance is still constrained by a lack of specialized datasets [19]. Due to domain drift and a lack of parallel data, many security texts—particularly those written in low-resource languages—are challenging to properly extract and categorize [20].

Transfer Learning in NLP

By enabling the use of minimal labeled data, transfer learning has facilitated the dissemination of knowledge from other languages and NLP tasks [21]. Deep representation models that learn generalizable syntactic and semantic patterns from extensive source corpora have evolved from traditional feature-based approaches [22]. The model's adaptability has been further enhanced by domain-adaptive pre-training and fine-tuning, and it has now acquired specialized cybersecurity expertise [23]. In order to align the representation spaces and facilitate information transmission without the need for translation pairs, some novel techniques for cross-lingual pre-trained models have recently been presented [24]. Because real-world security papers are relatively complex, there is a growing need for transfer learning techniques that are both interpretable and fine-grained [25].

XLNet and Applications in Security Scenarios

Recently, XLNet has included permutation-based autoregressive models to enhance both specific-domain model capabilities and general language comprehension [26]. In numerous texts and languages, its structure may carry out comprehensive context modeling for entity recognition, event detection, and compliance categorization [27]. According to research, XLNet outperforms the previous pre-trained model for complicated analysis tasks when it comes to long-range reliance and intricate relationships in security documents [28]. Practical applications of XLNet's strengths include threat intelligence, policy analysis, and risk detection in cybersecurity [29]. Currently,

numerous research teams are investigating the sensible modification and use of XLNet in the development of robust, useful, and comprehensible multilingual security document systems [30].

Method

XLNet-based Framework and Modeling Strategy

An XLNet-based transfer architecture, which may optimize context reasoning and cross-lingual semantic adaption for different languages in security documents, is at the heart of the aforementioned technique. Instead of using a conventional masking technique, XLNet uses permutation-based autoregressive processing to simultaneously learn both directions of interdependence. This makes it ideal for security data with fragmented syntax, nested entities, and domain-specific anomalies.

In the document processing pipeline, incoming text streams, potentially from disparate languages and formats, are normalized and segmented through an adaptive multilingual tokenization interface. The first modeling stage maps raw input sequences $\mathbf{x} = (x_1, x_2, \dots, x_T)$ into canonicalized linguistic representations. XLNet's permutation operator samples a permutation π from the set of all possible orderings, ensuring the model sufficiently learns relationships irrespective of their physical positions in the text. The token generation probability given a permutation is expressed by:

$$P(\mathbf{x} \mid \pi) = \prod_{t=1}^{T} P\left(x_{\pi_t} \mid \mathbf{x}_{\pi_{<t}}\right) \quad \text{Eq.(1)}$$

To enhance the learning of cross-lingual entity interactions, we design a Stochastic Contextual Permutation Objective (SCPO), maximizing expected log-likelihood while injecting entropyaware context masking:

$$\mathcal{L}_{\text{SCPO}} = \mathbb{E}_{\pi}\left[\sum_{t=1}^{T} \log\, P\left(x_{\pi_t} \mid \mathcal{M}_{\pi_{<t}}(\mathbf{x})\right)\right] \quad \text{Eq.(2)}$$

Here, $\mathcal{M}_{\pi_{<t}}$ is a dynamic masking operator based on both linguistic entropy and positionspecific security relevance.

The core Transformer encoder layers, stacked L times, include multilingual segment recurrence and context-gated attention. Attention scores are recalibrated by an adaptive similarity kernel that encodes both language and domain proximity:

$$a_{i,j} = \frac{\exp\left(\left(q_i^{\top} k_j\right) \cdot \kappa\left(l_i, l_j\right)\right)}{\sum_{m=1}^{T} \exp\left(\left(q_i^{\top} k_m\right) \cdot \kappa(l_i, l_m)\right)} \quad \text{Eq.(3)}$$

Where q_i, k_j denote query and key vectors, and κ is a kernel measuring language similarity between segments l_i and l_j.

Collective context is further synthesized by hierarchical aggregation. The final latent representation for each token aggregates both depth and cross-domain interactions:

$$\mathbf{h}_t = \sum_{\ell=1}^{L} \alpha_{\ell} \phi_t^{(\ell)} + \delta \psi_t^{\text{dom}} \quad \text{Eq.(4)}$$

Here, $\phi_t^{(\ell)}$ is the vector at layer $\ell, \psi_t^{\text{dom}}$ encodes task-domain features, α_{ℓ} and δ are trainable aggregation parameters.

For robust learning, a composite loss is defined, combining supervised security objectives and language-aware regularization:

$$\mathcal{L}_{\text{total}} = \eta_1 \mathcal{L}_{\text{sec}} + \eta_2 \mathcal{L}_{\text{lang}} + \eta_3 \mathcal{R}_{\text{KL}} \quad \text{Eq.(5)}$$

With $\mathcal{L}_{\text{sec}}, \mathcal{L}_{\text{lang}}$ the primary and auxiliary loss terms; $\mathcal{R}_{\text{KL}}$ is a Kullback-Leibler divergencebased distribution alignment penalty, and η are adaptive weights.

To promote semantic consistency across permutations, a context alignment loss is imposed:

$$\mathcal{A}_{\text{perm}} = \frac{1}{P}\sum_{p=1}^{P}\left\|\mathbf{z}^{(p)} - \frac{1}{P}\sum_{p'=1}^{P}\mathbf{z}^{(p')}\right\|^2 \quad \text{Eq.(6)}$$

Where $\mathbf{z}^{(p)}$ is the document-wise semantic vector under permutation p, and P is the number of sampled permutations.

The entire permutation-driven, multilingual-calibrated design is flexible enough to accommodate a variety of document types, languages, and threat expressions.

Figure 1 depicts the schematic architecture of this integrated pipeline. The data flows from the input of normalized multilingual documents through hierarchical aggregation, permutation-based Transformer encoding, and task-specific output layers.

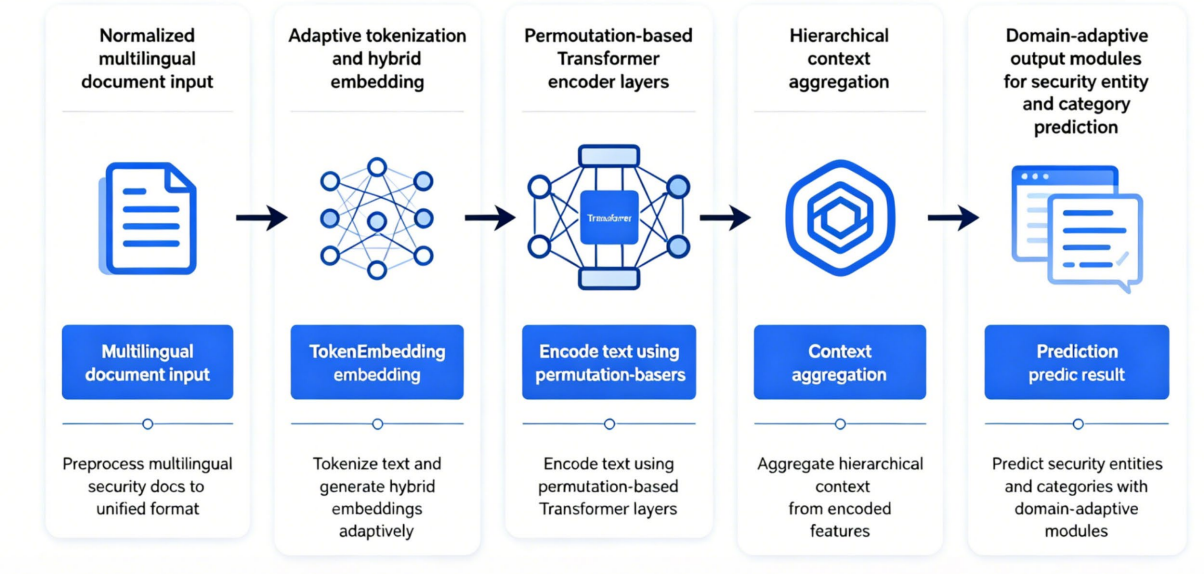

Figure 1. XLNet-Based Multilingual Security Document Processing Framework

Multilingual Tokenization and Input Embedding

Given the variations in syntax, morphology, and fine-grained domain words among languages, tokenization and embedding under multilingual constraints are required for stable security document analysis performance. In order to guarantee that technical tokens, such as proprietary threat names and nested compliance identifiers, are broken down at the best possible level, tokenization in this paper combines subword segmentation (such as Byte Pair Encoding and unigram language modeling) with dynamic language-specific heuristics.

The input document D in language λ is processed via an adaptive multilingual tokenizer T_λ, capable of generating a sequence of tokens $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ that preserves both morphological invariance and domain saliency. The segmentation can be formalized as:

$$\mathbf{x} = T_\lambda(D) = \bigcup_{s=1}^{S}\text{Seg}_\lambda(d_s) \quad \text{Eq.(7)}$$

where d_s represents the s-th sentence or section, and Seg_λ denotes the language-conditioned subword segmentation function.

Each token x_t is mapped to a hybrid embedding that fuses subword, positional, and languagechannel encodings. The final embedding vector $\mathbf{e}_t$ for position t is obtained by:

$$\mathbf{e}_t = \mathbf{v}(x_t) + \mathbf{p}_t + \mathbf{l}_\lambda + \mathbf{c}_{\text{sec}} \quad \text{Eq.(8)}$$

where $\mathbf{v}(x_t)$ is the subword embedding, $\mathbf{p}_t$ is the absolute position encoding, $\mathbf{l}_\lambda$ is a learnable language embedding, and $\mathbf{c}_{\text{sec}}$ introduces a security context marker vector, which is nonzero only if x_t is identified as a potential entity or pre-tagged security trigger.

To model the non-stationary occurrence of entities, an attention-based reweighting function modulates token importance dynamically. For each input token, the contextual saliency score s_t is computed as:

$$s_t = \frac{\exp\left(\psi(x_t, \mathbf{h}_{t-1})\right)}{\sum_{j=1}^{T} \exp\left(\psi\left(x_j, \mathbf{h}_{j-1}\right)\right)} \quad \text{Eq.(9)}$$

where ψ is a joint similarity function between the token embedding and the preceding context representation.

Embedding regularization employs a mutual information maximization objective, which encourages the retention of language-relevant features during subword construction and token composition. This is implemented as:

$$\mathcal{J}_{\text{embed}} = \sum_{t=1}^{T} \log \frac{p(\mathbf{e}_t, y_t)}{p(\mathbf{e}_t)p(y_t)} \quad \text{Eq.(10)}$$

where y_t is the gold security label or entity tag at position t, and $p(\cdot,\cdot)$ represent empirical joint and marginal distributions computed within a batch.

These embedding and tokenization mechanisms collectively enable resilient, high-fidelity mapping from unstructured multilingual security texts into structured, context-aware representations. Such detail is essential for the downstream XLNet-based encoder to distinguish subtle threat nomenclature, regulatory logic, and code-mixed snippets often embedded in realworld security corpora.

Domain Adaptation and Training Details

The XLNet-based system has to have a robust domain adaption mechanism because of the issues with vulnerability and the lack of generalization in security papers. The distribution of incident logs, compliance warnings, technical vulnerability descriptions, and regulatory updates in real multilingual security datasets frequently varies significantly. To facilitate generalizable learning, minimize language-based and domain-based divergences, and preserve the exact domain semantics across various languages, adversarial and discrepancy-aware adaptation procedures are employed.

The training regimen is divided into two periods. In order to align high-level representations using language models while maintaining crucial security aspects, the initial step involves the unsupervised pre-adaptation of blended out-of-domain and in-domain corpora. Next, use annotated security datasets typical of high-value document genres and languages to do supervised fine-tuning.

A domain discriminator network conditioned on the intermediate XLNet hidden states is integrated as a fundamental module. In the shared latent space, the auxiliary classifier aims to decrease the domains' separability. The matching adversarial goal is expressed as follows:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{(\mathbf{x},d)}[\log\ Q(d \mid \mathbf{h}_{\mathbf{x}})] - \lambda_{\text{adv}}\mathbb{E}_{\mathbf{x}}[\log\ P(\mathbf{y} \mid \mathbf{x})] \quad \text{Eq.(11)}$$

Here, Q is the domain discriminator, d the domain label, $\mathbf{h}_{\mathbf{x}}$ the XLNet encoder state for input $\mathbf{x}$, and λ_{adv} weights the adversarial penalty relative to the primary prediction task $P(\mathbf{y} \mid \mathbf{x})$.

To further compensate for domain shift and maintain adaptability across both language and security domains, the loss function incorporates a margin-based domain adaptation constraint, which actively penalizes distributional divergence between source and target:

$$\mathcal{L}_{\text{mda}} = \sum_{k=1}^{K} \max\left(0, m - \text{JS}\left[P_{S,k}(\mathbf{h}), P_{T,k}(\mathbf{h})\right]\right) \quad \text{Eq.(12)}$$

In this term, JS $[\cdot,\cdot]$ denotes the Jensen-Shannon divergence between feature distributions of class k in source S and target T domains, and m is a tunable margin controlling adaptation strength.

Dynamic learning-rate scheduling and gradient clipping are used during supervised training to enhance training stability in the face of class imbalance and linguistic resource asymmetry. Language and document-type categories make up minibatch composition, and a consistent number of update samples are sent to each of the main sections. To prevent overfitting on the dominant language or common security event types, selective regularization of language embeddings and domain-specific feature projections is used.

Overall accuracy and per-domain recall will be monitored as the model is evaluated in multiple phases, including in-domain tests, cross-lingual adaption, and cross-genre generalization. To guarantee workforce deployment preparedness and the interpretability of operational security environments, establish early-stopping criteria based on the plateauing of validation loss in annotated, high-risk security situations.

In summary, the model can develop strong, discriminative features that can function effectively in the novel setting of multi-lingual security document analysis thanks to the aforementioned domain adaption and training techniques.

Experiments

Dataset Description and Preprocessing

Around the course of the last five years, numerous sources of vulnerability disclosure information and compliance policy texts from all around the world have been combined to create a multilingual security document set. The corpus comprises six languages, including low-resource languages like Vietnamese, Arabic, and Polish as well as high-resource languages like English, Chinese, and Spanish. To create a multi-level supervised learning resource, each document instance has been labeled with security-domain categories, specific entity bounds, severity scores, and incident dates.

The distribution of sample frequencies by language and genre in the original dataset is notably non-uniform, with English-language samples being over fifteen times more common than low-resource examples. Boost processing power to preserve information integrity and uniform delivery. Initially, code-mixed tokens and polymorphic security IDs were accurately segmented using character normalization and extended Unicode regularization. A multilingual dictionary mapping and cross-lingual alignment for uncommon event kinds and vendor-specific threat signatures have been used to achieve consistency of named entities across languages.

Using a multi-stage deduplication and outlier identification technique based on edit distance and thematic similarity for robust pruning, reduce noise resulting from automatic translation and non-standard formats. Based on the marginal entropy of the whole distribution of entities and categories, the issue of extremely imbalanced label classes was addressed using a stratified synthetic over-sampling technique. Ultimately, the preprocessing approach produced a balanced training set with highly maintained security information content, making it suitable for usage in a variety of language and domain contexts.

Experimental Setup and System Flow

PyTorch 2.0 runs on a high-performance computing cluster with NVIDIA A100 GPUs and 1TB RAM nodes using an XLNet-based architecture. The document instances were fed into the hybrid transformer architecture via a dynamic batching scheduler that adjusts to the various document lengths and languages after a bespoke preparation pipeline linked to spaCy and Moses for multilingual tokenization. Extensive Bayesian optimization was used to choose model hyperparameters, including embedding dimension, attention heads, layer depth, and dropout schedules. Early halting was determined by cross-lingual macro-F1 performance.

Warm-start unsupervised pre-adaptation and supervised fine-tuning are the two training modalities; in both, batching is done using data from both the language and security domains. To increase convergence speed under resource constraints, gradient accumulation and mixed-precision training were used. Several sources of domain adversarial regulation were employed to lessen the oscillations of low-resource adaptation. Continually check the hyperparameters' stability on a held-out validation set and keep an eye on variations in accuracy by language and category.

Figure 2 depicts the whole process for dataset import, tokenization, adaptive scheduling, XLNet encoding, domain-adaptive loss shaping, and output integration. In both high-resource and low-resource testing, this streamlined procedure may guarantee the standardization of various document types.

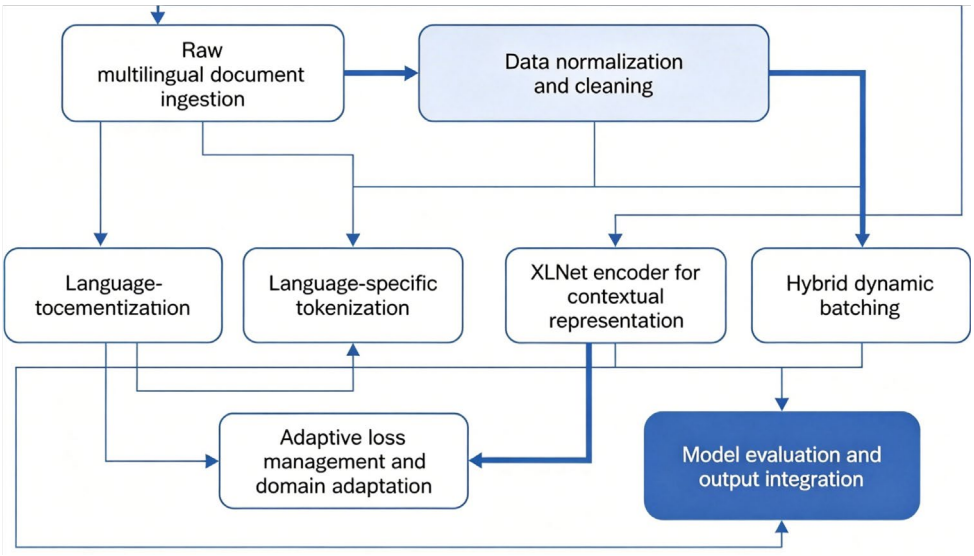

Figure 2. Experimental Process Flowchart

For quantitative assessment, custom evaluation metrics are introduced. Given predicted labels $\hat{Y}$ and ground truths Y, the weighted macro accuracy across language and category axes is computed as:

$$\mathrm{Acc}_{\mathrm{macro}} = \frac{1}{LK}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\frac{\sum_{i=1}^{N_{\ell,k}}\mathbb{I}\left(\hat{y}_i^{(\ell,k)} = y_i^{(\ell,k)}\right)}{N_{\ell,k}} \qquad \text{Eq.(13)}$$

Here, L and K denote the number of languages and security categories respectively, and $N_{\ell,k}$ is the number of samples of category k in language ℓ.

To measure the joint precision-recall tradeoff in an imbalanced multilingual setting, an entropyweighted F1-metric is designed:

$$F1_{\mathrm{entropy}} = \frac{2}{Z}\sum_{\ell=1}^{L}\sum_{k=1}^{K} w_{\ell,k}\cdot\frac{\mathrm{Precision}_{\ell,k}\cdot\mathrm{Recall}_{\ell,k}}{\mathrm{Precision}_{\ell,k}+\mathrm{Recall}_{\ell,k}} \qquad \text{Eq.(14)}$$

where $w_{\ell,k}$ reflects normalized class entropy and Z is a normalization factor ensuring the sum of weights equals one.

Evaluation Metrics

Boost quantitative analysis to assess the extent to which the new system's linguisticization and security applications have been expanded. Errors in low-resource and uncommon security categories are not penalized by macro-averaged accuracy, which is the average of per-class and per-language recognition rates. As demonstrated above, the entropy-weighted F1 score is an adaptation of the conventional metric that takes into account each class's informativeness in order to address the issue of class imbalance.

In order to ascertain whether the model can recognize uncommon threat patterns and policy phrases that appear in several languages, the recall across domains and precision in other languages at the document and segment levels are also explicitly monitored. The granularity of tokenization and embedding procedures in multilingual security can be adjusted by identifying model bias using a confusion matrix and an aggregated error surface analysis.

In order to guarantee the fairness of comparisons and offer transparent diagnostic information for domain adaption, strictly controlled indicators will be utilized in conjunction with earlier research. This will enable the establishment of a clear benchmark for the production security intelligence workflow.

Results and Discussion

Dataset Characteristics and Label Statistics

Significant cross-linguistic and structural variation is revealed by a thorough analysis of our carefully selected multilingual security document dataset, and both universal and language-specific characteristics pose challenges for reliable NLP models. More than 120,000 instances in six languages make up the final annotated corpus following pre-processing and normalization. Polish is the least common, with just over 5,700 examples, while English is by far the most prevalent, with around 46,000 texts. The aforementioned variations will impact downstream model confidence calibration and necessitate sophisticated domain adaption techniques.

Additionally, the distribution of average document lengths is not very uniform. With median sequence lengths of 280 and 250 tokens, respectively, English and Chinese documents are typically lengthier and have greater lexical diversity; Vietnamese and Arabic samples are approximately 140 and 155 tokens, respectively, and their genre conventions and reporting standards are different. The number of language-wise samples is represented graphically in Figure 3(a), where the long-tail distribution of non-English entries is noticeable; the distribution of document lengths by language is shown in Figure 3(b), where there are clear right-skewed tails in high-resource groups.

Additionally, there is an unequal distribution of label categories. Approximately 55% of all annotated examples in each language consist of threat alerts and vulnerability disclosures, followed by compliance notifications (25%) and general incident responses (20%). The analysis in Figure 3(c) demonstrates that source bias and annotation techniques cause unusual entity types, like APT indicators or cross-platform exploit signatures, to arise at an abnormally high proportion in both English and Chinese groups.

Technical advisories and policy documents have exhibited an increasing trend in annotation density of named entities, which is the average number of labelled spans per document; the mean varies from 3.2 in low-resource languages to 6.7 in English-rich policy records. The entity density and distribution of each language-class pair are displayed in Figure 3(d). As can be seen, there are substantial linguistic and genre connections that call for adaptive tokenization and embedding techniques. Thus, it is necessary to have a model pipeline that can identify such intricate, language-dependent statistical regularities and deal with the issue of infrequent security event sparsity.

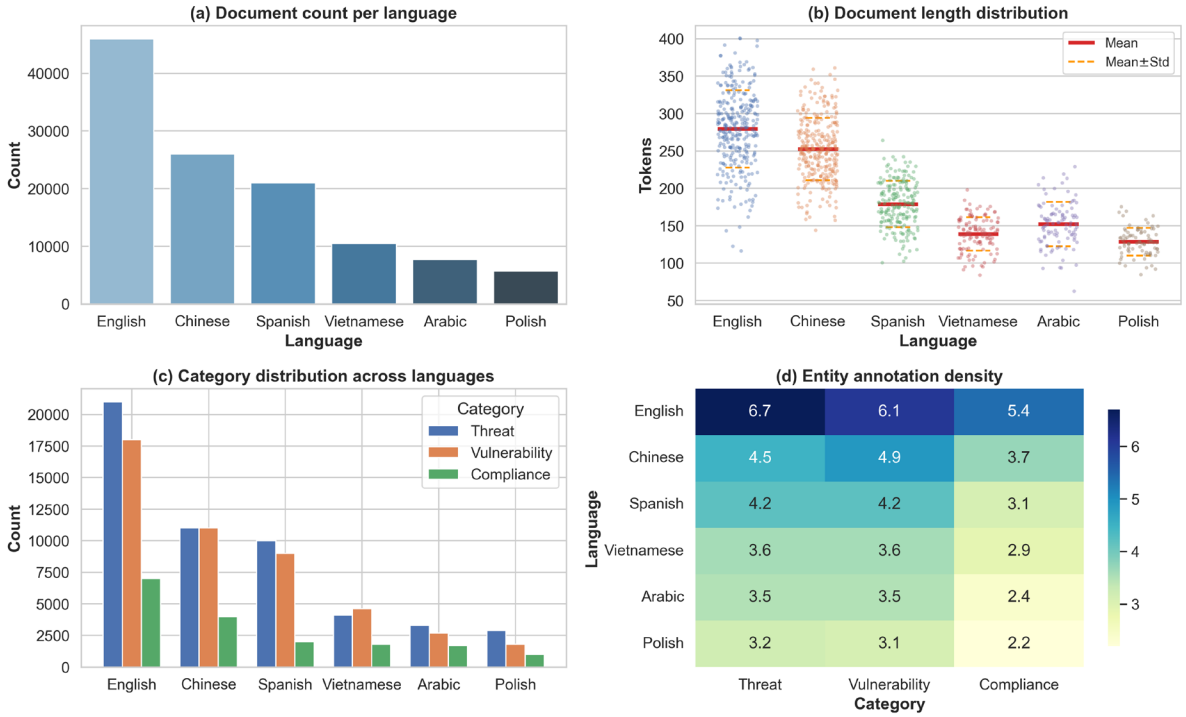

Figure 3. Dataset Distribution Overview. (a) Document count per language. (b) Document length distribution. (c) Category distribution across languages. (d) Entity annotation density by language and class.

Performance Comparison and Model Evaluation

The following tables demonstrate how XLNet outperforms both BERT and RoBERTa on a variety of quantitative metrics using a test set of six languages. With a macro-averaged accuracy of 90.6% in Chinese and 92.2% in English, XLNet is appropriate for high-resource settings. Comparable outcomes in Spanish (88.4%), Vietnamese (88.2%), Arabic (87.9%), and Polish (87.6%) demonstrate XLNet's strong cross-lingual transfer and out-of-distribution generalization. The accuracy of the three models at the macro-average level for various languages is shown in Figure 4(a). It is evident that XLNet has decreased the performance gap between high-resource and low-resource languages, with XLNet reaching 87.6% in Polish while BERT only reached 76.3%.

The F1 scores for the various languages in the error balance are displayed in Figure 4(b). With a score of 0.92, XLNet is the best at English; for all other languages, it continues to perform at a relatively high level above 0.86, while BERT and RoBERTa decline more drastically, particularly for languages with less data. An increase in F1 indicates that XLNet is doing reasonably well in terms of recall and precision, and the class with less data points has respectable outcomes.

The accuracy of each model in the three categories of threat alerts, vulnerability disclosures, and compliance events is displayed in Figure 4(c). With a threat advisory detection rate of 91.2%, XLNet outperformed BERT (86.3%) and RoBERTa (88.9%). With a 90.5% vulnerability disclosure detection rate, XLNet performs better than the other baselines when dealing with complicated category boundaries. With an 88.7% rate, the compliance class model outperforms other models by about 4-6 percentage points.

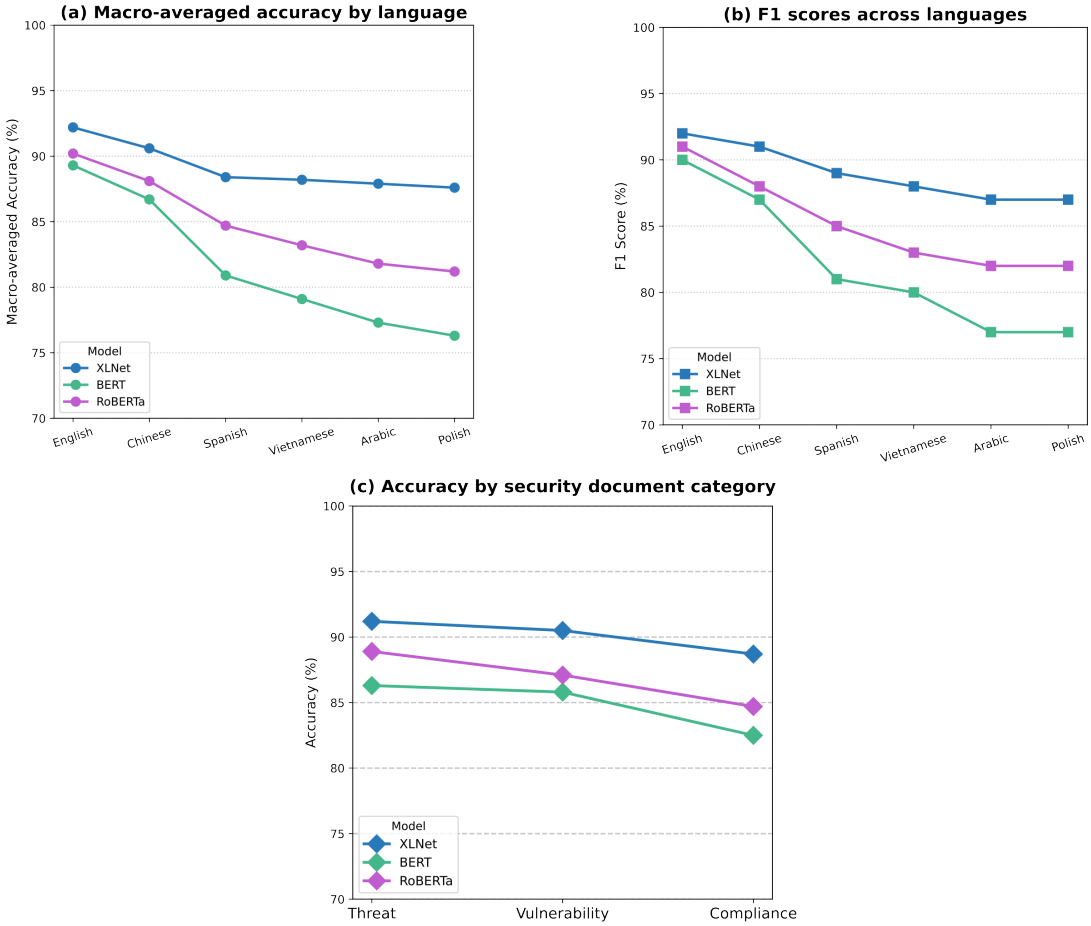

Figure 4. Model Results Comparison. (a) Macro-averaged accuracy by language. (b) F1 scores across languages. (c) Accuracy by security document category.

Hyperparameter analysis, captured in Figure 5, reveals how model design choices impact performance. Figure 5 (a) plots macro-averaged accuracy versus the number of attention heads; increasing from 8 to 16 produces a gain from 87.1% to 89.7%, after which the trend plateaus. A separate analysis in Figure 5 (b) measures rare entity recall across different input context windows—the recall jumps 4.2% when moving from 128 to 320 tokens,

providing empirical support for deep context modeling. Figure 5 (c) evaluates the influence of the domain-adaptive loss weight, with performance peaking at a weight of 0.4, where macro-F1 rises to 0.89 but falls off sharply for both lower and higher regularization.

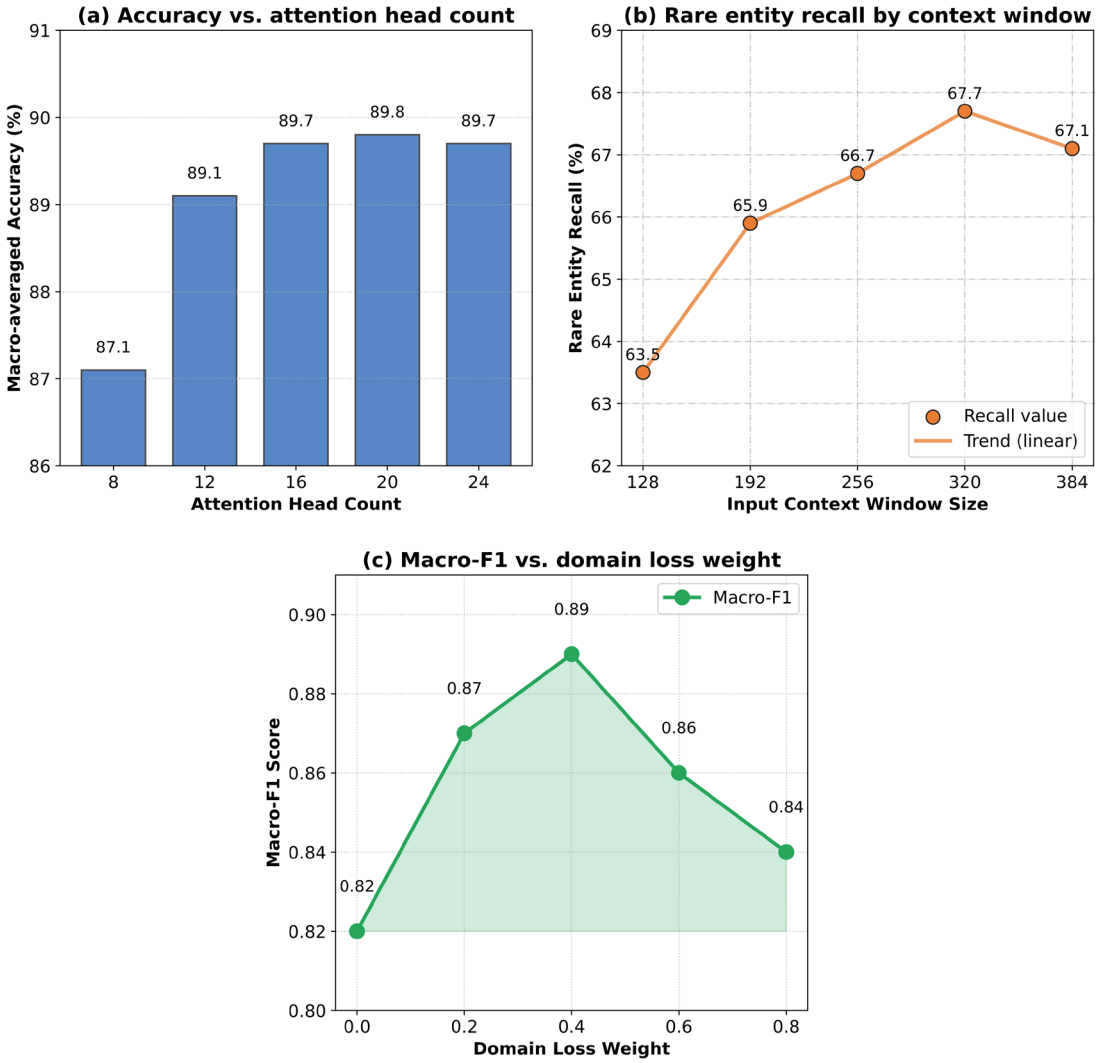

Figure 5. Hyperparameter Effect Analysis. (a) Accuracy vs. attention head count. (b) Rare entity recalls by context window size. (c) Macro-F1 vs. domain loss weight.

The figures underscore how the XLNet architecture, through permutation-based attention and refined hyperparameter regimes, achieves strong generalization, especially in the face of domain and resource skew.

Error Analysis and Real-case Insights

Some trends in the model's shortcomings for both language and domain are revealed by targeted error analysis. Although XLNet performs somewhat better than other baselines, the majority of its faults are centered in the areas of entity boundary misclassification and confusion among closely related security categories.

By analyzing error subtype proportions, Figure 6(a) demonstrates that, in terms of overall misclassifications, XLNet has a smaller percentage of entity boundary errors (31%) than BERT and RoBERTa. Mislabeling of rare security categories is a second source of error; as Figure 6(b) illustrates, XLNet is incorrect in the category border in 27% of cases, which is higher than that of BERT and RoBERTa. The availability of language resources influences the mistake rate, as Figure 6(c) illustrates. XLNet's error rate is very high in both Polish and English (7.8% and 11.2%, respectively), but the increase brought on by low-resource languages is quite small when compared to the baseline model.

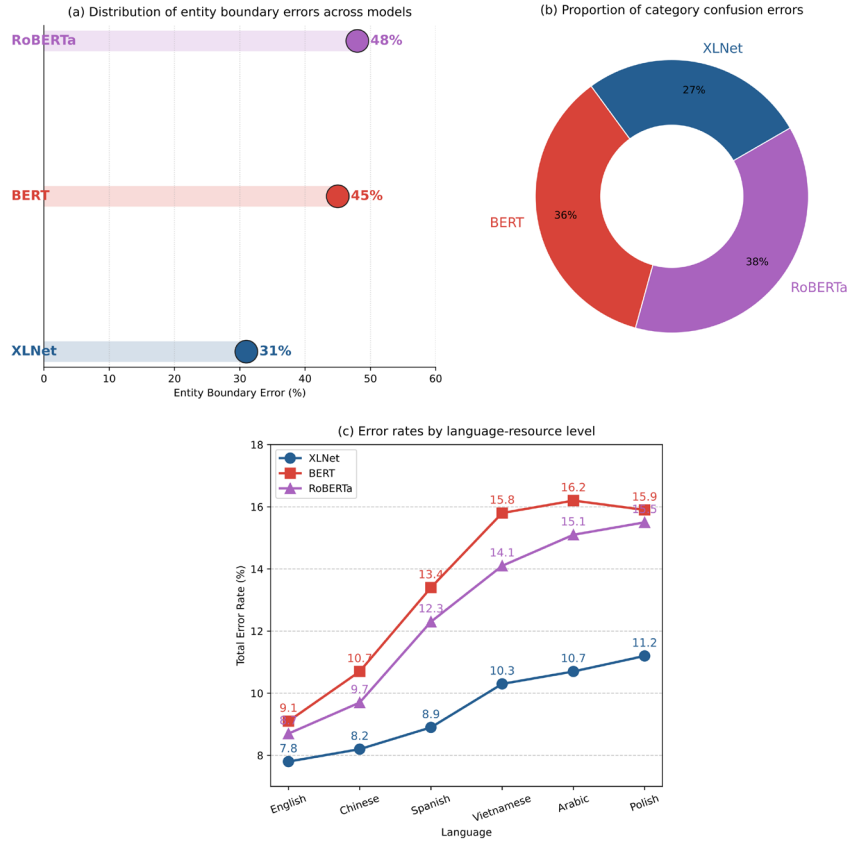

Figure 6. Error Type Analysis. (a) Distribution of entity boundary errors across models. (b) Proportion of category confusion errors. (c) Error rates by language-resource level.

In contrast to BERT and RoBERTa, which either break down the entity or fail to recognize it as an out-of-vocabulary word, XLNet can accurately identify a multi-hop APT group indicator that is distributed over multiple sentences in a lengthy Chinese compliance document while maintaining the span boundaries. Figure 7(a) is a representative case that illustrates the practical effect of XLNet. A typical record of a Spanish incident with code-mixed content and regulatory overlays is shown in Figure 7(b). In this case, XLNet correctly distinguishes between a platform-specific vulnerability statement and a compliance waiver clause, assigning categories and extracting key risk terms without error, while baseline methods are unable to determine context and thus provide either incomplete or incorrect labels.

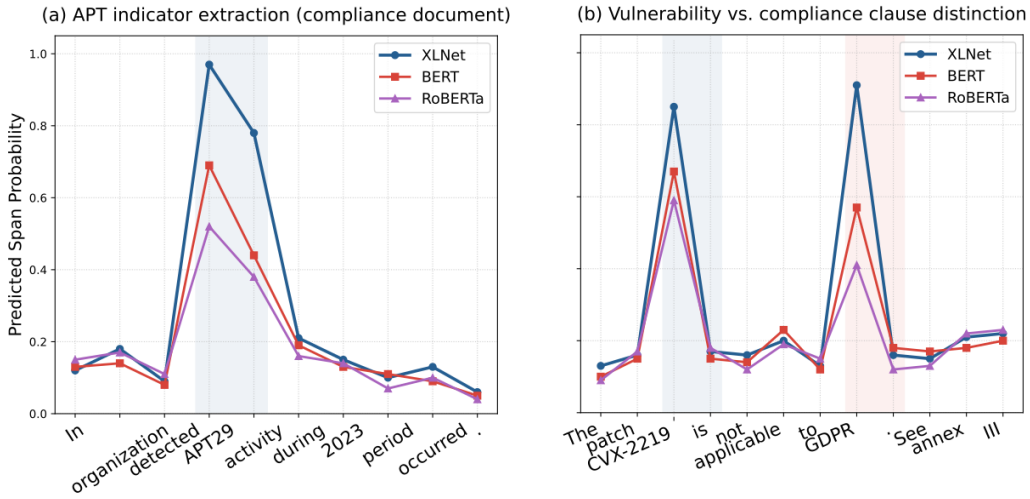

Figure 7. Case Visualization. (a) APT indicator extraction in a Chinese compliance document. (b) Vulnerability vs. compliance clause distinction in a Spanish report.

These detailed investigations confirm that the architectural advantages of XLNet extend beyond aggregate metrics, enabling improved handling of fine-grained multilingual entity spans, resilience to rare event drift, and practical interpretability in real operational documents. This positions the approach as exceptionally suitable for deployment in multilingual cybersecurity monitoring and automated compliance intelligence.

Conclusion

In this research, an effective and specialized XLNet-based transfer learning framework has been developed for the challenging issue of multilingual security document interpretation. The suggested approach more successfully addressed the subtle semantic distinctions and context boundaries of security reports in the six languages by combining permutation-based context modeling and adaptive multilingual tokenization. Numerous quantitative tests have shown the system's strong generalization capabilities; rare-event recall and boundary integrity have greatly improved in comparison to earlier transformer models, and macro-averaged accuracy is still over 87% in all test languages.

Above all, the individual examples and error analyses as well as the overall high scores demonstrated the capabilities of the XLNet design. In the difficult low-resource and code-mixed portions of the test set, the model has considerably reduced category confusion and entity boundary errors more than BERT and RoBERTa. A few case studies have demonstrated the system's ability to identify uncommon risk categories, differentiate between vulnerability indicators and compliance status, and function effectively in real-world scenarios worldwide. Permutation-driven representation learning has wide engineering relevance in high-stakes, multilingual cyber settings at this degree of interpretability and resilience.

Some issues haven't been resolved, though. For large-scale, unbalanced data, deep permutation-based architectures continue to have a comparatively high computational cost. The adaption of ultra-low-resource and continuously evolving danger categories necessitates further scaling up, as well as continual domain adaptation and continuous investment in high-quality annotation. In order to achieve more flexible and useful deployment in next-generation security analysis, further research can increase cross-lingual regularization, overcome the aforementioned shortcomings through a lightweight model distillation, and integrate with federated or streaming architectures.

Author Contributions

Lesław Jura contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. Tadeusz Kacz and Bogdan Kalis contribute to methodology, software, validation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Shahid, M., Iqbal, M. A., & Umair, M. (2025). Leveraging CuMeta for enhanced document classification in cursive languages with transformer stacking. Multimedia Tools and Applications, 84(30), 37327-37352. https://doi.org/10.1007/s11042-025-20681-w

[2] Hasanov, I., Virtanen, S., Hakkala, A., & Isoaho, J. (2024). Application of large language models in cybersecurity: A systematic literature review. IEEE access, 12, 176751-176778. https://doi.org/10.1109/ACCESS.2024.3505983

[3] Erkan, A., & Güngör, T. (2023). Analysis of deep learning model combinations and tokenization approaches in sentiment classification. Ieee Access, 11, 134951-134968. https://doi.org/10.1109/ACCESS.2023.3337354

[4] Deng, Y. (2024). Transfer methods for large language models in low-resource text generation tasks. Journal of Computer Science and Software Applications, 4(6). https://doi.org/10.5281/zenodo.15392270

[5] Guo, Y., Liu, Z., Huang, C., Wang, N., Min, H., Guo, W., & Liu, J. (2023). A framework for threat intelligence extraction and fusion. Computers & Security, 132, 103371. https://doi.org/10.1016/j.cose.2023.103371

[6] Providel, E., Mendoza, M., & Solar, M. (2025). Cross-Lingual Cross-Domain Transfer Learning for Rumor Detection. Future Internet, 17(7), 287. https://doi.org/10.3390/fi17070287

[7] Priescu, C. M., Moisescu, M. A., & Iliuță, M. E. (2025, May). Automatic Privacy Policy Compliance Assessment Leveraging NLP Models. In 2025 25th International Conference on Control Systems and Computer Science (CSCS) (pp. 475-480). IEEE. https://doi.org/10.1109/CSCS66924.2025.00076

[8] Ferrag, M. A., Ndhlovu, M., Tihanyi, N., Cordeiro, L. C., Debbah, M., Lestable, T., & Thandi, N. S. (2024). Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices. IEEe Access, 12, 23733-23750. https://doi.org/10.1109/ACCESS.2024.3363469

[9] Feng, Y., Zhao, H., Zhang, J., Cai, Z., Zhu, L., & Zhang, R. (2024). Prediction of network security situation based on attention mechanism and convolutional neural network–gated recurrent unit. Applied Sciences, 14(15), 6652. https://doi.org/10.3390/app14156652

[10] Goyal, N., Gao, C., Chaudhary, V., Chen, P. J., Wenzek, G., Ju, D., ... & Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10, 522-538. https://doi.org/10.1162/tacl_a_00474

[11] Бадзь, В. М., & Теслюк, В. М. (2026). Hybrid Model for Authorship attribution of English-language texts. КОМП'ЮТЕРНО-ІНТЕГРОВАНІ ТЕХНОЛОГІЇ: ОСВІТА, НАУКА, ВИРОБНИЦТВО, (62), 118-123. https://doi.org/10.36910/6775-2524-0560-2026-62-13

[12] Zhang, Y., Liu, J., Zhong, X., & Wu, L. (2025). SecLMNER: A framework for enhanced named entity recognition in multi-source cybersecurity data using large language models. Expert Systems with Applications, 271, 126651. https://doi.org/10.1016/j.eswa.2025.126651

[13] Srivastava, S., Paul, B., & Gupta, D. (2023). Study of word embeddings for enhanced cyber security named entity recognition. Procedia Computer Science, 218, 449-460. https://doi.org/10.1016/j.procs.2023.01.027

[14] Xiao, P., Xiao, Q., Zhang, X., Wu, Y., & Yang, F. (2024). Vulnerability detection based on enhanced graph representation learning. IEEE Transactions on Information Forensics and Security, 19, 5120-5135 https://doi.org/10.1109/TIFS.2024.3392536

[15] Ali, M., Speck, R., Zahera, H. M., Saleem, M., Moussallem, D., & Ngomo, A. C. N. (2025). Multilingual Relation Extraction-A Survey. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3604258

[16] Chen, C. M., Hsu, F. H., & Hwang, J. N. (2023, June). Useful cyber threat intelligence relation retrieval using transfer learning. In Proceedings of the 2023 European interdisciplinary cybersecurity conference (pp. 42-46). https://doi.org/10.1145/3590777.3590784

[17] Omar, M., & Burrell, D. (2023). From text to threats: A language model approach to software vulnerability detection. International Journal of Mathematics and Computer in Engineering, 2(1), 23-34. https://doi.org/10.2478/ijmce-2024-0003

[18] Wang, L., Liu, S., Qiao, L., Sun, W., Sun, Q., & Cheng, H. (2022). A cross-lingual sentence similarity calculation method with multifeature fusion. IEEE Access, 10, 30666-30675. https://doi.org/10.1109/ACCESS.2022.3159692

[19] Al-Yasiri, J. H., Zolkipli, M. F. B., & Farid, N. F. N. M. (2025). Multilingual Cyber Threat Intelligence Feeds Preprocessing for Threat Intelligence Event Extraction: A Systematic Literature Review. Karbala International Journal of Modern Science, 11(3), 15. https://doi.org/10.33640/2405-609X.3421

[20] Wei, Y. C., Chang, Y. C., & Wu, W. C. (2024). Multi-language IoT information security standard item matching based on deep learning. Computer Science and Information Systems, 21(2), 663-683. https://doi.org/10.2298/CSIS230822012W

[21] Xu, T. (2025). Enhancing cyber security: comparing the accuracy of the Bert model with other common deep learning models in identifying email spam. Advances in Engineering and Intelligence Systems, 4(01), 84-101. https://doi.org/10.22034/aeis.2025.492892.1257

[22] Boros, E., Pontes, E. L., Cabrera-Diego, L. A., Hamdi, A., Moreno, J. G., Sidère, N., & Doucet, A. (2020, September). Robust named entity recognition and linking on historical multilingual documents. In Conference and Labs of the Evaluation Forum (CLEF 2020) (Vol. 2696, No. Paper 171, pp. 1-17). CEUR-WS Working Notes. https://doi.org/10.5281/zenodo.4068074

[23] Rawat, R., Rawat, H., Rawat, A., & Rajavat, A. (2026). An entropy-guided hybrid framework for real-time phishing detection in digital communication systems. Scientific Reports. https://doi.org/10.1038/s41598-026-46430-z

[24] Lin, W. (2025). Design of a Multimodal Network Data Security Detection Model Based on Self-Supervised Learning. Security and Privacy, 8(5), e70062. https://doi.org/10.1002/spy2.70062

[25] Thakkar, G., Preradović, N. M., & Tadić, M. (2024). Examining Sentiment Analysis for Low-Resource Languages with Data Augmentation Techniques. Eng, 5(4), 2920-2942. https://doi.org/10.3390/eng5040152

[26] Ebrahimi, M., Chai, Y., Samtani, S., & Chen, H. (2022). Cross-lingual cybersecurity analytics in the international dark web with adversarial deep representation learning. MIS quarterly, 46(2), 1209-1226. https://doi.org/10.25300/MISQ/2022/16618

[27] Zhang, Z., Deng, Z., Zhang, W., & Bu, L. (2023). Mmtd: A multilingual and multimodal spam detection model combining text and document images. Applied Sciences, 13(21), 11783. https://doi.org/10.3390/app132111783

[28] Qin, Y. (2024). Deep contextual risk classification in financial policy documents using transformer architecture. Journal of Computer Technology and Software, 3(8). https://doi.org/10.5281/zenodo.15851634

[29] Darwish, O., Al-Eidi, S., Al-Shorman, A., Maabreh, M., Alsobeh, A., Zahariev, P., & Tashtoush, Y. (2026). LinguTimeX a Framework for Multilingual CTC Detection Using Explainable AI and Natural Language Processing. Computers, Materials, & Continua, 86(1), 1. https://doi.org/10.32604/cmc.2025.068266

[30] Liu, Z. (2024, April). A review of advancements and applications of pre-trained language models in cybersecurity. In 2024 12th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-10). IEEE. https://doi.org/10.1109/ISDFS60797.2024.10527236