Available on: http://gjetr.ep-journals.org/ E-ISSN: 3051-3782, P-ISSN: 3051-3774

Global Journal of Engineering and Technology Review

Volume: 02 Issue: 09 September 2026

Natural Language Processing for Real-Time Semantic Correlation of Shared Cyber Threat Indicators in National Defense

Chukwunenye Amadi
Independent Researcher, USA

Abolaji Adebayo
Computacenter, USA

Ayokunle Olamide Ijagbemi
Independent Researcher, USA

Published Date: September 05, 2026 **Article DOI : 10.65150/EP-gjetr/V2E9/2026-05**

ABSTRACT: Defense intelligence systems receive continuous streams of natural-language threat reports in which the same event is frequently described using different terminology, structure, and coded language. Traditional keyword systems classify such reports as unrelated, fragmenting the intelligence picture and delaying coordinated response. This paper examines how Natural Language Processing (NLP) enables the real-time semantic correlation of shared cyber threat indicators. It reviews the progression from term-weighting and word embeddings to contextual transformer representations, and specifies an NLP pipeline that combines semantic encoding, similarity scoring, named-entity and relationship extraction, and clustering to link related threats into unified, analyst-ready groups. Using representative open-source and synthetic corpora, the study shows conceptually and through illustrative results that embedding-based correlation recognizes semantically equivalent reports that share few surface tokens, thereby exposing coordinated and evolving threats. The paper contributes a structured account of semantic correlation for defense CTI, a categorization of similarity thresholds into actionable intelligence bands, and a discussion of the operational and ethical conditions for deployment. The evaluation is illustrative rather than comparative: the behavior described and the similarity bands reported are indicative defaults presented together with a specified evaluation protocol and calibration procedure, not measured results, and the relationship of this work to the first author's earlier study of semantic correlation and de-duplication is disclosed and delineated explicitly. Subject to those qualifications, the findings suggest that NLP-driven correlation materially improves situational awareness and the speed of national defense response.

KEYWORDS: natural language processing, semantic similarity, word embeddings, cyber threat intelligence, entity extraction, clustering, situational awareness.

I. INTRODUCTION

The defense of national infrastructure against sophisticated adversaries increasingly depends less upon the collection of raw data than upon the timely interpretation of the enormous textual record that surrounds every intrusion. Threat reporting arrives as a heterogeneous stream of advisories, incident narratives, malware analyses, and partner bulletins, each authored by a different organization with its own vocabulary, granularity, and assumed context. The operational challenge is not scarcity but correlation: distinguishing the moment at which two independently authored reports describe the same campaign, actor, or technique, so that fragmentary observations can be fused into a coherent operational picture. This capacity to relate disparate signals sits at the heart of situation awareness, the perception, comprehension, and projection of events within a volume of time and space that governs effective command decisions (Endsley, 1995). When correlation lags behind the tempo of an adversary, the defender operates on a stale representation of reality, and the value of shared intelligence decays accordingly. The problem is compounded by the volume of material, which routinely exceeds the finite processing capacity of human analysts working under time pressure (Miller, 1956). Contemporary surveys of cyber threat intelligence emphasize that the sharing of indicators has outpaced the ability of recipients to assimilate them, producing a paradox in which greater information exchange yields diminishing analytical returns unless accompanied by automated semantic reasoning (Tounsi & Rais, 2018; Wagner et al., 2019). This paper examines natural language processing as the enabling technology for that reasoning.

Framing correlation as an information-theoretic act clarifies what automated systems must accomplish. Each incoming report carries a quantity of information whose relevance depends entirely upon what the recipient already knows; a bulletin restating established facts contributes little, whereas one resolving an ambiguity carries disproportionate value (Shannon, 1948). The defensive analyst therefore acts as a channel of limited capacity attempting to extract mutual information between newly arrived text and an existing model of adversary behavior. Structured intelligence exchange formats were introduced precisely to reduce ambiguity in this channel, encoding indicators, relationships, and observables in machine-readable schemas (Barnum, 2014), while behavioral taxonomies provide a shared vocabulary of adversary tactics and techniques against which

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

observations may be aligned (Strom et al., 2018). Yet the majority of actionable knowledge continues to originate as unstructured prose, written for human consumption and never fully translated into these schemas. The intelligence value locked in narrative text remains inaccessible to systems that reason only over structured fields. Bridging this gap requires representations that capture meaning rather than surface form, and correlation procedures that operate over those representations at the speed of the incoming stream. The remainder of this work argues that recent advances in semantic text representation make such real-time correlation tractable, and situates that argument within the operational constraints peculiar to national defense.

The defense context imposes requirements that distinguish it from commercial text analytics. Adversaries are adaptive and adversarial in the literal sense: they deliberately vary terminology, employ euphemism and code words, and structure their operations to frustrate pattern matching. The kill chain model describes intrusions as a sequence of dependent stages, from reconnaissance through actions on objectives, and correlation across reports frequently means recognizing that two accounts describe different stages of a single campaign rather than unrelated events (Hutchins et al., 2011). Such recognition cannot rest on shared keywords, because the language describing reconnaissance bears little lexical resemblance to the language describing exfiltration even when both concern the same actor. Furthermore, defense intelligence tolerates neither the latency of batch processing nor the error profile of opaque automation; analysts require systems whose associations are auditable and whose false correlations are rare, because a spurious link can misdirect scarce resources during an active incident. These pressures motivate a careful treatment of how meaning is encoded and compared. The surveys that map this rapidly maturing field consistently identify semantic correlation and cross-source fusion as the principal open problems, noting that extraction techniques have advanced faster than the methods for relating what is extracted (Sun et al., 2023). This paper concentrates on that relational layer, treating extraction as a prerequisite rather than an endpoint.

The contribution of this first half of the paper is a structured account of the representational and correlational machinery required for real-time semantic fusion of shared threat indicators, organised to expose where existing techniques suffice and where the defense setting demands more. It proceeds from the failure modes of naive lexical matching, through the evolution of text representation from sparse term weighting to contextual and sentence-level embedding, to the specific tasks of indicator extraction, topic discovery, and clustering that convert prose into correlatable structure. Throughout, the discussion foregrounds the tension between semantic fidelity and computational tractability, since a correlation engine that cannot keep pace with the ingestion rate offers no operational advantage regardless of its accuracy. The treatment deliberately spans the boundary between the natural language processing literature, which has produced the representational advances, and the cyber threat intelligence literature, which has articulated the operational requirements but drawn less consistently on those advances. By reading the two together, the paper aims to make explicit the design choices facing anyone building a correlation system for national defense. The analysis assumes familiarity with neither field in depth, and each technical construct is introduced in terms of the correlation problem it addresses, so that the argument remains legible to the interdisciplinary audience that operational cyber defense necessarily comprises.

Because the problem addressed here has been examined previously by the first author, that earlier study is disclosed at the outset rather than left for a reader to discover. Amadi (2025) proposed and evaluated a framework applying large language models and natural language processing to the real-time semantic correlation and de-duplication of shared threat indicators, and reported that embedding-based correlation compared favorably with conventional keyword matching on open-source and synthetic intelligence data. The present paper is not an extension of that evaluation and advances no comparative performance claim of its own. Its object is the architecture, the calibration discipline, and the interpretive apparatus through which semantic correlation can be operated and audited in a coalition setting, and every quantity it reports is illustrative rather than measured. Section V.A states the relationship between the two studies in full, distinguishing them in method, data, results, and claimed contribution.

II. THE SEMANTIC CORRELATION PROBLEM IN DEFENSE INTELLIGENCE

A. Why Keyword Matching Fails

The intuitive first approach to relating threat reports is to compare the words they contain, and the classical information retrieval apparatus formalises this intuition. Documents are represented as vectors over a vocabulary, with each term weighted by a scheme that rewards local frequency while discounting terms common across the collection (Salton & Buckley, 1988). Two reports are then judged similar to the degree that their weighted term vectors point in the same direction. This machinery underpins search engines of remarkable effectiveness and remains a sensible baseline (Manning & Raghavan, 2008). Yet its central assumption, that shared meaning manifests as shared vocabulary, breaks down precisely in the cases that matter most to defense. The vocabulary mismatch problem is structural rather than incidental: distinct authors describe identical phenomena with disjoint word choices, so that a report referring to a compromised credential and another referring to a stolen password may share no discriminating term despite describing the same event. Term weighting cannot rescue such pairs, because the discounting of common terms actively suppresses the very general words that might otherwise bridge them, while the rare technical tokens that carry weight are exactly those most likely to diverge between authors. The consequence is systematically depressed recall on semantically equivalent but lexically divergent reports, an error mode that is invisible in aggregate retrieval benchmarks but corrosive to correlation, where the missed link is the whole point of the exercise.

Beyond vocabulary mismatch, lexical matching suffers from the twin pathologies of polysemy and synonymy that no purely surface representation can resolve. A single token frequently denotes unrelated concepts across contexts; a term naming a common utility may reference a benign administrative tool in one report and a weaponised living-off-the-land technique in another, and a keyword system conflates the two, manufacturing false correlations that waste analytic attention. Conversely, the many distinct expressions for a single adversary technique fragment its mentions across incompatible surface forms, scattering evidence that ought to accumulate. Sparse representations also discard word order and syntactic structure entirely, collapsing a report into an unordered bag whose meaning is only approximated by term co-occurrence. This is tolerable when documents are long and topically redundant, so that aggregate statistics dominate, but threat indicators are often terse, and the difference between an actor targeting a system and an actor originating from a system may hinge on a single preposition that the representation cannot see. The defense setting amplifies every one of these weaknesses because adversaries actively exploit them, and because the cost of a missed correlation is measured

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

in operational consequence rather than a marginal drop in a ranking metric. These limitations establish the requirement that motivates the remainder of the paper: representations that encode meaning rather than form.

B. Meaning, Context, and Coded Language

Meaning in natural language is contextual, and this fact is not a nuisance to be engineered away but the property a correlation system must exploit. The same word contributes different meaning depending on its neighbors, and human analysts resolve such ambiguity effortlessly by reading each token in light of the surrounding discourse. Early distributional approaches captured only a static, context-independent sense for each word, assigning one vector regardless of usage, which is inadequate when the discriminating signal lies in how a term is used rather than whether it appears. Representations that condition a word's encoding on its entire sentence changed this picture decisively, producing token representations that shift with context and thereby disambiguate polysemous terms automatically (Peters et al., 2018). Architectures that attend jointly over all positions in a passage carried the principle further, yielding encoders in which every token's representation is a function of the complete surrounding text (Devlin et al., 2019). For threat correlation this is transformative, because the meaning of a technical term in an incident narrative is precisely what determines whether two reports concern the same behavior. A representation that reads a compromised host mentioned as a pivot point differently from the same host named as an initial victim can preserve distinctions that surface matching erases, and can recognize equivalences that surface matching misses, provided the correlation procedure is built atop such contextual encodings rather than atop raw tokens.

National defense introduces a further layer that ordinary text analytics rarely confronts: language engineered to conceal. Adversaries and the communities that surround them employ code words, deliberate misspellings, evolving slang, and euphemism, so that the surface form of a report may be chosen specifically to defeat automated detection. Correlation must therefore be robust to substitutions that preserve meaning while destroying lexical identity, recognizing that a novel codename introduced in one channel refers to an actor already described under a different label elsewhere. This robustness cannot come from any fixed lexicon, which adversaries render obsolete as quickly as it is compiled, but must instead emerge from representations that place semantically related expressions near one another in a continuous space even when those expressions have never co-occurred. The requirement is doubly demanding because coded language often carries its meaning through context alone: a benign-seeming phrase acquires threat significance only from the operational setting in which it appears. A correlation system must consequently reason over meaning at the level of phrases and passages rather than isolated tokens, and must do so with sufficient generalisation to accommodate expressions absent from any training corpus. These demands frame the representational program that the next section develops, tracing how the field moved from counting words toward encoding the meanings they carry.

III. TEXT REPRESENTATION: FROM TERM WEIGHTING TO CONTEXTUAL EMBEDDINGS

A. Sparse and Distributional Representations

The representational history relevant to correlation begins with the sparse, high-dimensional vectors of classical retrieval, in which each dimension corresponds to a vocabulary term and most entries are zero. Weighting these dimensions by local and global frequency statistics yields representations that are interpretable, efficient, and surprisingly strong for topical matching (Salton & Buckley, 1988), and the accumulated engineering around them remains foundational to practical systems (Manning & Raghavan, 2008). Their defining limitation is that dimensions are orthogonal: distinct terms occupy independent axes, so the representation encodes no notion that two words are related in meaning, and semantic proximity is invisible by construction. The distributional hypothesis, that words appearing in similar contexts tend to carry similar meanings, motivated a shift toward dense low-dimensional vectors learned so that this hypothesis is reflected geometrically. Neural methods estimated such word vectors efficiently from large corpora by predicting words from their contexts or contexts from their words (Mikolov et al., 2013a), and refinements showed that the resulting spaces encoded regular semantic and syntactic relationships as consistent geometric offsets (Mikolov et al., 2013b). In these spaces, synonyms and related technical terms cluster together, directly addressing the vocabulary mismatch that defeats sparse matching. For the correlation problem this is the pivotal conceptual advance, because it makes semantic similarity a measurable quantity, a distance in a continuous space, rather than a binary question of shared tokens that terse and adversarially varied threat reports so often answer in the negative.

Distributional word embeddings admit several complementary formulations, each with consequences for defense applications. One influential approach derived vectors from global word co-occurrence statistics, factorising a corpus-wide matrix so that vector differences captured meaning ratios (Pennington et al., 2014), offering a bridge between the count-based tradition of classical retrieval and the predictive tradition of neural estimation. A distinct and operationally important refinement represented each word as a composition of character n-grams rather than an atomic unit, so that the embedding of an unseen word could be assembled from its subword components (Bojanowski et al., 2017). This subword capability matters acutely in the threat domain, where malware families, tool names, and adversary handles are coined continually and where deliberate misspelling is a routine evasion tactic; a model that can embed a never-before-seen token by its morphology degrades gracefully where a fixed vocabulary fails outright. Nonetheless, all of these methods share a decisive limitation for correlation: each word receives a single vector irrespective of the sentence in which it occurs, so polysemy remains unresolved and the context-dependent meanings that distinguish related threat behaviors are averaged into an indiscriminate whole. The representation of a passage as an aggregate of such static vectors, whether by simple averaging or by weighted composition, inherits this defect and additionally discards word order. These constraints established the agenda for contextual representation, in which a token's vector is computed afresh from its surroundings on every occurrence.

B. Contextual and Sentence-Level Encoders

The move to contextual representation resolved the static-vector limitation by making each token's encoding a function of its sentence. Deep bidirectional language models produced representations in which a word's vector reflects both left and right context, so that distinct senses of a term receive distinct encodings (Peters et al., 2018). The attention-based architecture that subsequently dominated the field dispensed with recurrence entirely, computing each position's representation by attending over all others in parallel, which improved both the modeling of long-range dependencies and the efficiency of training at scale (Vaswani et al., 2017). Pre-training a deep bidirectional encoder of this form on a masked-token objective yielded contextual representations of exceptional generality, transferable across many downstream tasks with minimal

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

adaptation (Devlin et al., 2019). Careful study of the pre-training regime demonstrated that longer training on more data with revised objectives improved these representations further (Liu et al., 2019), while a unified text-to-text formulation showed that a single encoder-decoder could subsume diverse tasks under one interface (Raffel et al., 2020). Very large autoregressive models later exhibited the capacity to perform tasks from instructions and few examples alone (Brown et al., 2020). For threat correlation, contextual encoders supply exactly the sense-disambiguating, context-sensitive token representations that coded and technical language demands, allowing a system to distinguish uses of a term that surface matching and static embeddings alike conflate, and to recognize equivalences that depend on operational context rather than shared vocabulary.

Correlation, however, operates over passages and reports rather than isolated tokens, and token-level contextual encoders do not by themselves yield a single vector whose geometry reflects sentence meaning. Deriving such sentence representations from a masked language encoder naively, by averaging token vectors or reading a designated summary position, produces embeddings whose distances correlate poorly with semantic similarity, which is fatal for a system that must judge whether two reports match by comparing vectors directly. Purpose-built sentence encoders addressed this by training on objectives that shape the geometry of the sentence space explicitly. One line supervised the encoder on inference data so that entailment and contradiction structured the space (Conneau et al., 2017); another trained a general-purpose encoder optimized for transfer across sentence-level tasks (Cer et al., 2018). A siamese adaptation of a pre-trained bidirectional encoder produced sentence embeddings comparable by simple distance, reducing the cost of pairwise comparison from an intractable joint encoding of every candidate pair to a single encoding per report followed by cheap vector arithmetic (Reimers & Gurevych, 2019). Contrastive objectives refined the space further, pulling paraphrases together and pushing unrelated sentences apart to sharpen similarity judgments (Gao et al., 2021). This efficiency is decisive for real-time defense: correlation over an incoming stream requires that each new report be embedded once and matched against an index by nearest-neighbor search, a design that the sentence-encoder paradigm makes feasible at operational scale.

IV. INFORMATION EXTRACTION AND CORRELATION FOR THREAT INTELLIGENCE

A. Indicator and Entity Extraction

Semantic representation supplies the substrate for correlation, but operational intelligence also requires that the specific entities and indicators within a report be identified and typed, so that correlation can be grounded in concrete observables as well as diffuse similarity. A substantial body of work has addressed the automatic extraction of security-relevant entities from unstructured text, developing labeled corpora and taggers that recognize actors, malware, vulnerabilities, and infrastructure within prose (Bridges et al., 2013). Complementary efforts framed the task as populating structured knowledge from security writing, extracting entities and their relationships to build queryable representations of threat knowledge (Joshi et al., 2013). Systems targeting indicators of compromise specifically learned to isolate the addresses, hashes, and domain names embedded in narrative reports, distinguishing genuine observables from incidental mentions (Liao et al., 2016). Extraction has also been pushed beyond atomic indicators toward the behavioral level: one influential system mined threat actions from analytical prose and mapped them onto adversary tactics and techniques, converting descriptive text into structured behavioral intelligence (Husari et al., 2017). This behavioral mapping aligns naturally with the standardised taxonomies of adversary technique (Strom et al., 2018) and with the structured exchange formats that carry indicators between organizations (Barnum, 2014), so that extraction becomes the bridge from human-authored narrative into machine-correlatable structure. The broader discipline situating these methods treats such extraction as an instance of data-driven security analytics, in which learning systems convert raw security text into actionable structured knowledge (Sarker et al., 2020).

Extraction from threat advisories and live feeds presents difficulties beyond those of general entity recognition, and the literature has confronted them directly. Advisories are written in dense, jargon-laden prose whose relevant facts are often expressed as high-level assertions rather than named entities, and dedicated systems have been built to extract structured intelligence from such documents at the level of the claims they make (Ramnani et al., 2017). Supervised methods have targeted the higher-order intelligence embedded in reports, extracting not merely indicators but the threat context that renders them actionable (Ghazi et al., 2018). The problem sharpens further when the source is social media, where reports are short, noisy, and time-critical, yet may carry the earliest warning of an emerging campaign; systems built for this setting monitor streams and raise alerts by extracting and assessing security-relevant mentions in near real time (Mittal et al., 2016). The recurring lesson across these efforts is that extraction alone does not accomplish correlation: it yields a set of typed observations that must still be related across reports, deduplicated, and fused into coherent entities. Extracted indicators anchor correlation by providing exact-match evidence where it exists, while semantic representation supplies the connective tissue where surface forms diverge. A correlation architecture for defense therefore layers the two, using extraction to establish grounded anchors and embeddings to relate the surrounding narrative, so that neither the precision of the former nor the recall of the latter is sacrificed.

B. Topic Modeling and Clustering

Correlation at scale ultimately requires grouping: relating not merely pairs of reports but organising an entire corpus into coherent themes, campaigns, and actor profiles that analysts can navigate. Probabilistic topic models offered an early unsupervised route, representing each document as a mixture over latent topics and each topic as a distribution over words, thereby uncovering thematic structure without supervision (Blei et al., 2003). Applied to threat reporting, such models can surface the dominant subjects across a feed and track their evolution, providing a coarse organising layer above the individual report. Clustering methods complement topic modeling by partitioning documents or their embeddings directly according to proximity, and the extensive literature on text clustering catalogues the algorithms and similarity measures suited to high-dimensional textual data (Aggarwal & Zhai, 2012). The choice of algorithm carries operational weight: methods requiring a predetermined number of clusters are ill-suited to an open, evolving threat landscape in which the number of active campaigns is unknown and changing, whereas density-based approaches discover clusters of arbitrary shape and explicitly label outliers as noise without fixing the cluster count in advance (Ester et al., 1996). This outlier-awareness is doubly valuable in defense, since a report that clusters with nothing may signal a genuinely novel campaign deserving of immediate attention rather than a mere anomaly to be discarded, and the ability to flag such novelty is itself an intelligence product.

The effectiveness of both topic modeling and clustering depends decisively on the representation over which they operate, which ties this section back to the representational program of the paper. Applied to sparse term vectors, clustering inherits the vocabulary-mismatch failures already

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

discussed, grouping reports by shared tokens rather than shared meaning, so that lexically divergent accounts of one campaign scatter across clusters. Applied instead to contextual sentence embeddings, the same algorithms group reports by semantic proximity, drawing together accounts that share meaning while differing in surface form, which is exactly the behavior correlation requires. This layering, contextual embedding followed by density-based clustering with explicit novelty detection, constitutes a practical architecture for organising a live threat stream into evolving campaign clusters that an analyst can monitor. Reviews of the field confirm both the promise and the immaturity of these pipelines, observing that while extraction from unstructured threat text has advanced considerably, the downstream fusion of extracted and embedded information into stable, interpretable groupings remains an active frontier (Rahman et al., 2020). The surveys mapping the discipline as a whole reach a consistent verdict: representation and extraction have matured faster than correlation and fusion, and the integration of semantic representation with scalable clustering under real-time constraints is where the most consequential open problems for national defense now lie (Sun et al., 2023).

V. RELATED WORK: APPLIED CYBER DEFENSE AND CROSS-SECTOR INTELLIGENT SYSTEMS

A. Relationship to the Authors' Prior Work

The correlation problem treated in this paper was examined previously by the first author. Amadi (2025) presented a framework applying large language models and natural language processing to the real-time semantic correlation and de-duplication of shared cyber threat indicators, applied it to open-source and synthetic intelligence datasets, and reported that the embedding-based approach compared favorably with conventional keyword matching in both accuracy and processing speed. The present manuscript addresses the same operational problem, and it would be misleading to present that framing as novel. What is new lies in the method it specifies, the data on which it relies, the results it claims, and the contribution it advances, and each of these is delineated below so that the boundary between the two studies is unambiguous.

In methodology, the earlier study is a system demonstration: it assembles an end-to-end pipeline and benchmarks it against a lexical baseline, with that comparison as the object of the work. The present paper is a specification and analysis. It decomposes correlation into five separable stages, namely normalization, semantic encoding, similarity scoring, entity and relationship extraction, and incremental clustering with consolidation; it fixes the fusion rule under which distributed and symbolic evidence jointly license a link rather than either acting alone; and it introduces a calibration and banding discipline governing how a continuous similarity signal is converted into differentiated operator action. Three architectural commitments in particular originate in this paper and are not carried over from the earlier evaluation: the concurrent rather than sequential execution of extraction and encoding in order to bound latency, the incremental maintenance of clusters so that an arriving indicator can extend, merge, or seed groups without reprocessing the window, and the treatment of the banded output as a decoding rule over a noisy channel rather than as a single decision boundary.

In data, the earlier study assembled open-source and synthetic intelligence datasets in order to support a comparative benchmark, and its measurements rest upon them. The present paper contributes no corpus, reuses no measurement from that study, and derives no conclusion from its datasets. The representative stream described in Section VII exists solely to exhibit qualitative regularities that follow from the design, and Section VII.A specifies in full the protocol, annotation scheme, metrics, and reporting requirements under which a substantive evaluation should instead be conducted.

In results, the earlier study reports comparative accuracy and processing speed against a keyword baseline. The present paper reports no comparative quantity and makes no claim of superiority over any alternative system. Its findings are the stratification of the similarity range into regimes carrying distinct operational meanings, the interpretation of that stratification through situation awareness, information theory, and the cognitive capacity of the analysts who must act on the output, and the identification of the conditions under which the stratification remains valid. The numeric boundaries in Table II are illustrative defaults rather than estimates, and the calibration procedure that would replace them is specified in Section VII.

In scientific contribution, the earlier study establishes feasibility, showing that semantic correlation of shared indicators is achievable and preferable to lexical matching. The present paper stands to it as an architectural and interpretive account stands to a feasibility demonstration. It contributes a structured reading of the representational and correlational machinery that the defense setting requires, a confidence-band framework that converts a continuous similarity signal into differentiated and auditable analyst dispositions, and an account of the operational, ethical, and governance conditions under which such a system may be fielded across a coalition. None of these contributions depends on the earlier findings, and the argument advanced here would stand unchanged were those findings revised.

B. Applied Cyber Defense and Cross-Sector Intelligent Systems

The techniques surveyed above do not exist in isolation; they have been assembled into applied systems across cyber defense and adjacent domains where large volumes of specialised text must be interpreted under time pressure, and where correlation across independently authored sources is the operative challenge. Situating the present contribution requires examining how prior systems have combined representation, extraction, and correlation into working pipelines, what architectural compromises they accepted to meet operational constraints, and where their design choices reveal the boundaries of current practice. It also benefits from looking beyond the security domain to other high-stakes settings, from clinical text analysis to financial surveillance to open-source intelligence, in which the same fundamental tension between semantic fidelity and real-time throughput has produced instructive solutions. Reading these cross-sector systems alongside the cyber defense literature exposes recurring patterns, common failure modes, and transferable design principles that a correlation engine for national defense would do well to inherit. The following section undertakes this comparative review, first surveying applied cyber defense platforms that operationalise the representational and extractive techniques already described, and then drawing selectively on intelligent systems from other sectors whose confrontation with scale, latency, and adversarial or noisy input illuminates the design space for the real-time semantic correlation architecture this paper ultimately proposes.

The framework advanced in this paper is positioned against a broad and rapidly growing body of applied scholarship. Three strands are especially pertinent: research on cyber defense and threat intelligence, which defines the operational problem; research on artificial intelligence and secure digital systems, which supplies the analytical machinery; and a wider literature that applies data-driven modeling, risk governance, and intelligent

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

decision support across other critical-infrastructure and enterprise sectors, which demonstrates the breadth and transferability of the methods on which real-time semantic processing draws. The remainder of this section surveys these strands in turn.

A mature stream of applied research addresses the detection, prioritization, and containment of evolving threats in enterprise and national settings. This work develops maturity models, threat-actor analysis, incident-response and security-orchestration frameworks, and threat-informed defense engineering that together define the operational requirements the present study seeks to accelerate: rapid triage of high-volume indicators, measurable control effectiveness, and coordinated response across organizational boundaries (Adebayo et al., 2023; Akeju et al., 2018; Dosunmu & Ogundele, 2021, 2022, 2023, 2024a, 2024b, 2024c, 2025a, 2025b; Ekechi et al., 2022; Mbonu et al., 2019a, 2021a; Odejobi et al., 2025; Ogbole et al., 2025; Smart, 2020, 2023; Smart & Jooda, 2023a, 2023b, 2025, 2026).

Complementary studies examine the analytical capabilities that artificial intelligence and machine learning bring to security and decision support, including predictive analytics, human-in-the-loop annotation, generative models, edge intelligence, and interpretable learning. Collectively they establish both the promise of automated inference over large, noisy datasets and the necessity of retaining human oversight where consequences are severe (Akanbi et al., 2025; Aliliele et al., 2023a; Annan, 2024; Asiedu et al., 2026; Borah et al., 2022; Dagodzo et al., 2022; Eboh & Aliliele, 2025a; Elebe et al., 2023; Essandoh et al., 2025; Jooda & Smart, 2024a; Komi, 2021, 2023; Komi & Adamolekun, 2021; Komi & Adeniji, 2023; Komi & Ganiu, 2023; Ladapo et al., 2022a, 2024, 2026; Mayo et al., 2021; Mbonu et al., 2021b; Okoruwa et al., 2025; Quainoo et al., 2025a; Tonoyan et al., 2021a, 2022a).

A substantial governance literature specifies the control, oversight, and accountability conditions under which intelligent systems must operate, spanning regulatory compliance, data-protection and privacy engineering, identity and access management, continuous auditing, and audit-automation. These contributions are directly relevant to fielding language models in defense, where transparency, provenance, and role-based control are prerequisites rather than afterthoughts (Adebayo et al., 2025; Aliliele et al., 2023b, 2024a, 2024b, 2025a, 2025b; Annan, 2022, 2025a, 2025b; Badmus et al., 2019a, 2022a, 2024, 2025; Bello et al., 2024; Dosunmu & Ogundele, 2025c; Edivri et al., 2026; Ekechi et al., 2026; Eyetsemitan et al., 2020; Fadayomi et al., 2021; Jooda & Smart, 2024b; Kumuyi et al., 2023, 2024; Mbonu et al., 2018, 2019b, 2020a, 2020b, 2020c, 2022a, 2022b, 2022c; Ogbole et al., 2021a; Okoruwa et al., 2020; Smart & Jooda, 2024a).

Work on cloud engineering, secure software delivery, and enterprise integration informs the engineering foundations required to deploy language-based analytics reliably at scale, addressing configuration governance, integration patterns, secure testing, and the operational discipline needed to keep data pipelines dependable under load (Aliliele et al., 2023c; Asiedu & Quainoo, 2023a, 2023b, 2025; Asiedu et al., 2025; Badmus et al., 2018, 2019b, 2022b; Eboh & Aliliele, 2023, 2024, 2025b; Ladapo et al., 2022b, 2023, 2025; Quainoo et al., 2024, 2025b; Tonoyan et al., 2021b).

Beyond the security domain, research on supply-chain optimization, procurement, and logistics demonstrates closely analogous data-driven approaches to coordinating, de-duplicating, and governing information across distributed, multi-stakeholder operations. The problems of reconciling heterogeneous records, eliminating redundant entries, and maintaining a single authoritative view mirror those confronted in threat-intelligence consolidation (Agbabiaka et al., 2019; Dosunmu & Ogundele, 2019, 2024d; Eyetsemitan et al., 2024; Ladapo et al., 2019; Ogunwole et al., 2021; Okonkwo et al., 2023, 2024a, 2024b, 2025; Okoruwa et al., 2024; Oteri & Edivri, 2024, 2026; Tonoyan et al., 2024a).

Parallel work on energy systems, smart grids, and renewable infrastructure applies predictive modeling, continuous monitoring, and resilience engineering to complex critical-infrastructure environments analogous to those defended here, offering transferable insights on real-time inference, degradation detection, and the equitable, dependable operation of systems under stress (Asiedu & Quainoo, 2024; Atta et al., 2025; Dagodzo, 2018a; Komi & Adamolekun, 2022; Komi & Adeniji, 2020; Komi & Ganiu, 2022).

Studies of industrial and occupational safety contribute mature models of hazard detection, near-miss signal analysis, and operational risk governance. Their central concern, distinguishing weak leading signals from background noise before an incident occurs, directly parallels the signal-versus-noise framing that motivates semantic filtering and prioritization in this study (Arumosoye & Obriki, 2019, 2021, 2023; Atta et al., 2020; Obogo et al., 2019, 2021, 2025; Obriki & Arumosoye, 2018, 2019, 2022, 2023; Obriki et al., 2023).

Research on healthcare and diagnostic infrastructure offers further evidence on lifecycle management, resilience planning, and performance measurement for mission-critical systems that must remain dependable under public-health stress, reinforcing the importance of governance, standardization, and measurable outcomes when scaling sensitive systems (Aminu-Ibrahim & Ogbete, 2023; Aminu-Ibrahim et al., 2025a, 2025b; Asiedu, 2023; Ogbete & Aminu-Ibrahim, 2024; Ogbete et al., 2023; Ozowara et al., 2025).

Contributions on geospatial intelligence, UAV and remote-sensing systems, and aviation-safety oversight demonstrate how heterogeneous sensor and spatial data are fused into actionable operational intelligence and how regulatory oversight is exercised over autonomous platforms, a fusion-and-oversight challenge closely related to correlating multi-source threat indicators (Adeyelu, 2023; Dagodzo, 2018b, 2021, 2022, 2025; Smart & Jooda, 2024b).

A body of work on financial analytics, fraud detection, and business-process optimization shows how analytical models detect anomalies, reduce redundancy, and streamline decision workflows in high-stakes, data-intensive settings, providing methodological parallels for the anomaly-sensitive, redundancy-averse processing pursued in defense intelligence (Edivri & Oteri, 2022; Mbonu et al., 2021c; Ozowara et al., 2022; Tonoyan et al., 2022b, 2023, 2024b, 2026).

Further contributions on data analytics, business intelligence, and continuous monitoring illustrate how meaning and actionable structure are extracted from large, heterogeneous data streams for timely decision making (Arumosoye et al., 2026; Obogo et al., 2026; Ogbole et al., 2021b).

VI. A PIPELINE FOR REAL-TIME SEMANTIC CORRELATION

A. Pipeline Overview

The proposed system organizes real-time semantic correlation of shared cyber threat indicators into a sequence of loosely coupled stages, each of which transforms an incoming stream of heterogeneous reports into progressively more structured and actionable representations. At the ingestion boundary, indicator submissions arriving from allied sensors, incident responders, and open-source feeds are normalized into a common schema that preserves the free-text narrative alongside any structured metadata. The narrative payload is then passed to a semantic encoding stage that

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

maps variable-length text into fixed-dimensional vectors suitable for rapid comparison. A similarity scoring stage evaluates each newly encoded indicator against a rolling window of recent submissions, surfacing candidate correlations that exceed calibrated thresholds. In parallel, an entity and relationship extraction stage isolates the atomic observables and tactical descriptors embedded in each narrative, providing a symbolic complement to the distributed vector representation. Downstream, a clustering and consolidation stage groups mutually reinforcing indicators into coherent incident hypotheses and suppresses redundant restatements of the same underlying event. The design deliberately separates encoding, scoring, extraction, and consolidation so that individual components can be upgraded, retrained, or replaced without disturbing the remainder of the pipeline. Table I summarizes the principal techniques associated with each stage, the representational form each produces, and the analytic role it serves within the broader objective of maintaining shared situational understanding across a distributed defensive coalition operating under stringent latency constraints.

Table I. NLP techniques and their role in semantic correlation

Technique	Function	Purpose in correlation
Contextual embeddings	Convert text into meaning vectors	Enable semantic comparison
Cosine similarity	Measure vector closeness	Quantify relatedness
Named-entity recognition	Identify actors, places, events	Build threat networks
Relationship extraction	Link entities across reports	Reveal coordinated activity
Clustering	Group related reports	Consolidate the intelligence picture

B. Semantic Encoding

The semantic encoding stage constitutes the representational foundation of the pipeline, because the fidelity with which free-text indicators are projected into vector space determines the ceiling of every subsequent correlation decision. Contextual language models based on the transformer architecture provide bidirectional representations that capture the polysemy and syntactic dependence characteristic of technical security prose (Devlin et al., 2019). However, the token-level embeddings such models emit are not directly comparable at the level of whole reports, since naive pooling degrades the geometric coherence needed for meaningful distance computation. Sentence-level encoders address this deficiency by fine-tuning a shared network with a siamese objective, producing embeddings whose cosine geometry aligns with human judgments of semantic equivalence (Reimers & Gurevych, 2019). Complementary encoders trained on large heterogeneous corpora with multitask objectives offer robust general-purpose representations that transfer to specialized domains with modest adaptation (Cer et al., 2018). More recent contrastive approaches demonstrate that even unlabeled text can yield discriminative sentence embeddings by contrasting an instance against dropout-perturbed variants of itself, which is attractive in the defense setting where labeled correlation pairs are scarce and sensitive (Gao et al., 2021). Within the pipeline, the encoder is treated as an interchangeable module: reports are batched, encoded once, and cached, so that the amortized cost of representation is incurred a single time per indicator and reused across all pairwise comparisons, extraction cross-checks, and clustering operations performed downstream, which is decisive for meeting the throughput demands of a coalition feed.

C. Similarity Scoring

Given fixed-dimensional embeddings, the similarity scoring stage quantifies the semantic proximity between a newly ingested indicator and the population of recently observed reports. Cosine similarity serves as the primary measure, as it is invariant to embedding magnitude and emphasizes directional agreement in the learned representation space, which corresponds most closely to shared meaning. To satisfy real-time constraints, exhaustive pairwise comparison is avoided in favor of approximate nearest-neighbor retrieval over an index that is incrementally maintained as the rolling window advances. Each query returns a bounded candidate set whose scores are then subjected to threshold calibration; rather than adopting a single global cutoff, the system maintains adaptive thresholds conditioned on the density of the local neighborhood, so that a match asserted in a sparse region of the space carries the same evidentiary weight as one asserted in a dense region. Scores are not consumed as binary decisions but are retained as continuous weights that propagate into the consolidation stage, where they contribute to the strength of hypothesized links between indicators. This continuous treatment preserves uncertainty until sufficient corroborating evidence accumulates, and it allows the operator interface to rank correlations by confidence rather than presenting an undifferentiated set of asserted equivalences. The scoring stage additionally records the temporal offset between correlated indicators, enabling downstream reasoning about the propagation of an observed campaign across the coalition and the sequence in which partners first observed a shared threat.

D. Entity and Relationship Extraction

Distributed vector representations excel at capturing holistic semantic similarity, yet they deliberately abstract away the discrete observables that analysts must ultimately act upon, such as addresses, file hashes, domains, and named adversary tooling. The entity and relationship extraction stage recovers this symbolic structure by applying sequence-labeling models trained to recognize security-relevant entities within technical narratives. Foundational work on security entity extraction established that domain-adapted labeling substantially outperforms generic named-entity recognizers, because the lexical and orthographic conventions of indicators diverge sharply from those of ordinary prose (Bridges et al., 2013). Building on isolated entity recognition, the pipeline also extracts the tactical behaviors described in each report, mapping verb-centric descriptions of adversary actions onto a structured vocabulary of techniques. Automated extraction of threat actions from unstructured text has been shown to bridge the gap between prose reporting and machine-actionable behavioral descriptions, allowing narratives to be indexed by what an adversary did rather than merely by which artifacts were observed (Husari et al., 2017). Within the correlation objective, these extracted behaviors provide a second, orthogonal axis of comparison: two reports may reference entirely distinct observables while describing an identical sequence of tactics, and such behavioral congruence is frequently more durable than artifact overlap, since adversaries rotate infrastructure far more readily than they redesign their operational playbooks. The stage therefore emits both an entity set and a behavioral signature for every indicator it processes, and both are carried forward as first-class evidence.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

The extracted entities and behaviors are assembled into a lightweight relationship graph that records which observables co-occur within a report and how those observables connect to the tactical steps attributed to the adversary. This graph is fused with the similarity scores from the preceding stage so that correlation decisions rest on converging evidence from both distributed and symbolic representations, mitigating the failure modes each incurs in isolation. Purely lexical matching of indicators is brittle against paraphrase and obfuscation, whereas purely semantic matching can conflate superficially similar but operationally distinct reports; the fusion of the two provides mutual disambiguation. The design draws on systems that convert streaming public discourse into structured, timely alerts by coupling entity recognition with relevance filtering, demonstrating that symbolic extraction and rapid dissemination are compatible objectives even at social-media velocity (Mittal et al., 2016). To bound latency, extraction is executed concurrently with encoding rather than sequentially, and the resulting graph fragments are merged into a persistent knowledge structure that accumulates across the rolling window. Entities are canonicalized through defanging normalization and alias resolution so that syntactic variants of the same observable collapse to a single node. The persistent graph subsequently serves as both a corroboration source for ambiguous semantic matches and an audit trail, allowing an analyst to trace precisely why the system proposed a given correlation and which extracted elements substantiated it, which is essential to sustaining trust in an automated coalition workflow.

E. Clustering and Consolidation

Once indicators have been scored and enriched with symbolic structure, the clustering and consolidation stage aggregates individually correlated pairs into coherent groups that correspond to distinct threat events or campaigns. Surveys of text clustering establish that the choice of representation and distance function dominates cluster quality, which motivates operating directly on the semantic embeddings rather than on sparse term-frequency vectors that would fragment paraphrased reports (Aggarwal & Zhai, 2012). Because the number of distinct campaigns present in a stream is unknown a priori and fluctuates continuously, partitioning methods that require a predetermined cluster count are ill-suited to the streaming setting. Density-based clustering offers a more appropriate inductive bias, as it discovers clusters of arbitrary shape, requires no advance specification of their quantity, and explicitly designates low-density points as noise (Ester et al., 1996). This noise-aware property is valuable operationally, since a large fraction of ingested indicators are isolated observations that should not be forced into any campaign until corroborating evidence arrives. The stage applies density-based grouping over the fused similarity structure, treating the continuous scores as a proximity graph and promoting a group to a candidate incident only when its internal cohesion and size exceed configured minima. Clusters are maintained incrementally so that an arriving indicator can extend, merge, or seed groups without reprocessing the entire window, which is essential to preserving the real-time responsiveness the operational context demands of the system even as the reporting volume grows.

Beyond grouping, the consolidation stage produces a compact, human-consumable summary of each cluster so that operators confront a small number of synthesized incident hypotheses rather than an unbounded feed of raw submissions. Probabilistic topic modeling contributes to this synthesis by inferring latent thematic structure across the reports assigned to a cluster, yielding descriptive term distributions that characterize the campaign without manual labeling (Blei et al., 2003). These inferred themes are combined with the most representative extracted entities and behaviors to form a concise incident profile, together with provenance links back to the constituent reports and their originating partners. Redundancy suppression operates at this level as well: when multiple partners independently report the same event, the consolidation logic recognizes the semantic and symbolic overlap and collapses the restatements into a single hypothesis annotated with corroboration count and source diversity, both of which serve as confidence signals. The stage also tracks the temporal evolution of each cluster, so that a campaign that reactivates after a period of dormancy is linked to its prior manifestation rather than being registered as a novel event. By emitting consolidated, provenance-rich, confidence-weighted hypotheses at a cadence matched to operator cognition, the stage converts the high-volume, high-redundancy input stream into a manageable set of decision-relevant objects, which is the concrete artifact through which the pipeline delivers shared understanding to the coalition it serves and the point at which machine processing yields to human judgment.

VII. ILLUSTRATIVE RESULTS AND INTERPRETATION

A. Evaluation Protocol and the Status of Reported Quantities

The epistemic status of everything reported in this section should be stated plainly before the observations are described. No controlled experiment is claimed. The corpus is representative rather than benchmarked, the arrival rates are configured rather than sampled from an operational feed, and there is no annotated ground truth against which precision and recall could be computed. Every quantity appearing in this section and in Table II is therefore illustrative: it indicates the shape of the behavior the design is expected to exhibit and the ranges within which practitioners commonly operate, and it carries no evidential weight in support of the design's superiority over any alternative. The claims of this paper rest on the architectural argument of Section VI and the interpretive framework developed below, not on measured performance, and the band boundaries in Table II should be treated as defaults to be replaced by calibration before any operational use.

For that reason the protocol under which a substantive evaluation should be conducted is specified here in sufficient detail to be executed and audited. The corpus would consist of shared indicator reports drawn from open advisory feeds and partner bulletins, augmented with controlled restatements: paraphrases that preserve the referent while varying vocabulary and structure, machine translation and back-translation to reproduce multilingual restatement, and reports describing distinct kill-chain phases of a single campaign. Ground truth would be defined at two levels, a pairwise relation over reports distinguishing the same event, the same campaign at a different phase, and unrelated activity, and a partition of reports into campaigns. Both would be established by at least two independent annotators, with inter-annotator agreement reported as a chance-corrected coefficient and the adjudication of disagreements documented. Negative pairs would deliberately include hard negatives, that is reports sharing generic tradecraft language but no common referent, since these are precisely the cases in which semantic proximity is ambiguous and against which the fusion rule must be tested. Corpus size, the proportion of positive pairs, the number of distinct campaigns, the language distribution, and the temporal span would all be disclosed, because the reduction in reviewable volume is trivially inflated by a corpus with a high redundancy rate and cannot be interpreted unless that rate is known.

Three families of measurement follow from the three outcomes the pipeline is intended to affect. Linking quality would be reported as precision, recall, and their harmonic mean over the pairwise relation, computed across the full range of candidate thresholds and summarized as a precision-

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

recall curve with average precision, rather than at a single operating point that would conceal the very trade-off the bands exist to expose. Grouping quality would be reported against the campaign partition using measures that tolerate an unknown cluster count and an explicit noise class, such as B-cubed precision and recall alongside the adjusted Rand index, together with the proportion of reports assigned to noise, which is itself an operationally meaningful quantity. Consolidation would be reported as the ratio of distinct objects presented to the analyst after consolidation to distinct objects ingested, paired with the fraction of genuine campaigns that survive consolidation as separate hypotheses, since volume reduction achieved by over-merging is a harm rather than a benefit and the two figures are uninterpretable in isolation. Latency would be reported as a distribution rather than a mean, giving the median and the ninety-fifth and ninety-ninth percentiles of end-to-end time from ingestion to hypothesis emission at each arrival rate, alongside sustained throughput and the queue depth at which the real-time guarantee fails.

The configuration producing any such figures would be reported in full: the encoder and its version, the sequence-length truncation, the similarity function, the density parameters governing neighborhood radius and minimum cluster size, the window over which clusters are maintained, the extraction models, and the normalization and alias-resolution rules, since correlation quality depends on these choices at least as strongly as on the choice of algorithm family. Attributing observed behavior to particular components requires ablation rather than end-to-end comparison alone, so a lexical baseline, embeddings without symbolic fusion, symbolic extraction without embeddings, and the full fused configuration would each be evaluated over the identical corpus and partition. Because clustering outcomes vary with stream order and initialization, each configuration would be run repeatedly over shuffled arrival orders, with results reported as means accompanied by bootstrap confidence intervals rather than as single values, and differences between configurations tested rather than asserted. Reporting of this form is what would convert the qualitative regularities described below into evidence; in its absence they remain what they are here, expectations grounded in the design and in the prior literature.

B. Observed Regularities and Their Interpretation

To characterize the behavior of the pipeline under conditions representative of coalition threat sharing, an illustrative characterization was constructed over a representative stream of indicator reports exhibiting the paraphrase, redundancy, and multilingual restatement patterns typical of distributed defensive reporting. The characterization emphasizes interpretation over competitive benchmarking, since the contribution of this work lies in articulating how semantic correlation reshapes the analyst's information environment rather than in advancing a single accuracy metric. Reports are taken to arrive at controlled rates, and the pipeline is characterized with respect to three families of outcome: the proportion of genuinely related reports it successfully linked, the proportion of unrelated reports it erroneously linked, and the reduction in the volume of distinct objects an analyst would need to review after consolidation. Across arrival rates spanning ordinary and surge conditions, the expected behavior is that the system links semantically equivalent reports sharing no lexical overlap, since distributed representations recover correspondences that are invisible to string matching. Erroneous links should be expected to concentrate among reports describing generic activity lacking distinctive observables or behaviors, where semantic proximity is genuinely ambiguous. The consolidation stage should produce a substantial contraction in reviewable volume, collapsing redundant restatements into a smaller set of corroborated hypotheses. These qualitative regularities, rather than any single headline figure, constitute the object of the present characterization and motivate the banded interpretation developed in the remainder of this section as a disciplined way of reading the system's continuous confidence output.

A structural feature of the similarity signal is that correlation quality is not uniform across the similarity range but instead stratifies naturally into regimes with distinct operational meanings. At the upper extreme of the similarity scale, links are dominated by near-duplicate restatements and confirmed equivalences, and precision is expected to approach saturation. In an intermediate regime, links reflect genuine but partial relatedness, such as reports describing different phases of a common campaign or overlapping subsets of infrastructure, where the fusion of semantic and symbolic evidence is decisive in separating true relationships from coincidental proximity. At the lower extreme, similarity signals become unreliable and are better treated as weak priors that require corroboration before being surfaced. Recognizing these regimes allows the system to attach differentiated handling and differentiated analyst guidance to each, rather than imposing a single decision boundary that would either flood operators with speculative links or discard subtle but important correspondences. Table II formalizes this stratification into interpretive confidence bands, associating each band with its characteristic similarity regime, the dominant type of relationship it captures, the corroboration it warrants, and the recommended disposition for an operator. The banded framing is intentionally representational rather than prescriptive, since the exact numeric boundaries depend on the encoder, the domain, and the tolerance for false linkage that a given coalition is prepared to accept in exchange for broader recall across its shared reporting, and they must be recalibrated whenever any of those conditions changes materially.

Table II. Illustrative interpretation of semantic similarity scores. The boundaries are indicative defaults requiring calibration, not estimated values.

Similarity band	Score range	Interpretation
Near-identical meaning	0.85 - 0.95	Very high semantic match
Related threats	0.70 - 0.84	Moderate semantic link
Weakly related	0.50 - 0.69	Partial similarity
Unrelated	Below 0.50	No semantic correlation

The boundaries in Table II are indicative defaults consistent with common practice for sentence-level encoders, not values estimated from the stream described above, and they should be replaced by calibrated equivalents before deployment. Calibration proceeds from the annotated development set described in Section VII.A. For a chosen encoder and corpus, the pairwise precision-recall curve is computed, and the upper boundary is set at the smallest similarity at which precision meets the target the coalition sets for automatic consolidation, commonly a value close to unity, because an incorrect merge silently destroys a distinct hypothesis and leaves no trace for an analyst to recover. The lower boundary is set at the similarity below which the rate of true relations is statistically indistinguishable from the base rate of the negative sample, so that scores beneath it are treated as carrying no evidence rather than weak evidence. The intermediate boundary separating links that warrant corroboration from those that warrant analyst review is chosen against the coalition's review capacity rather than by a statistical criterion, since it determines the

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

volume routed to human judgment; it is the one boundary that should be set by operational rather than empirical argument. Each boundary should be reported with an uncertainty interval obtained by bootstrapping the development set, recalibrated whenever the encoder, the language mix, or the reporting population changes materially, and monitored in deployment through periodic re-annotation of a sample of surfaced links, since drift in the reporting stream will move the operating point even when the model is held fixed.

The interpretive value of these bands is best understood through the lens of situation awareness, which distinguishes the perception of elements in the environment, the comprehension of their meaning, and the projection of their future status (Endsley, 1995). The high-confidence band operates primarily at the level of perception and comprehension, confirming that dispersed reports refer to the same underlying event and thereby establishing a coherent picture from fragmentary observations. The intermediate band contributes most directly to projection, since linking distinct phases of a campaign allows analysts to anticipate an adversary's trajectory rather than merely cataloguing discrete artifacts. Viewed information-theoretically, the consolidation achieved by the pipeline can be read as a reduction in the entropy of the analyst's input distribution, concentrating probability mass onto a smaller number of distinguishable incident hypotheses (Shannon, 1948). Each redundant restatement that the system collapses carries little marginal information once its equivalent has been observed, and suppressing it raises the average information content of the objects that ultimately reach the operator. The correlation links themselves act as a channel that transmits the structure latent in the raw stream, and the banded confidence framework functions as a decoding rule that manages the trade-off between the errors of asserting spurious links and omitting genuine ones, a trade-off that is inherent to any inference performed over a noisy shared reporting environment and that no threshold can eliminate, only position advantageously.

The consolidation behavior also bears directly on the cognitive constraints of the human operators the system is intended to support. Classic work on the capacity of human information processing established that individuals can hold only a small number of distinct items in immediate attention before performance degrades (Miller, 1956). An unconsolidated stream of coalition indicators routinely presents analysts with a volume of distinct objects that exceeds this capacity by orders of magnitude, guaranteeing that some correlations will be missed simply because no individual can attend to every pairwise relationship across a high-velocity feed. By contracting the stream into a bounded set of corroborated hypotheses, each accompanied by provenance and a confidence band, the pipeline aligns the quantity of decision-relevant objects with the range that human attention can accommodate. The banded presentation further supports selective attention, permitting operators to triage by confidence and to reserve scarce analytic effort for the intermediate regime, where machine judgment is least certain and human expertise is most valuable. In this sense the quantitative reduction in reviewable volume is not merely an efficiency gain but a prerequisite for the shared understanding the coalition seeks, since situational awareness that exceeds the cognitive capacity of those who must act upon it confers no operational benefit. The results thus connect the technical behavior of the pipeline to the human ceiling it is ultimately designed to respect rather than to overwhelm.

VIII. DISCUSSION

The foregoing pipeline and its illustrative behavior situate semantic correlation within a broader trajectory of cyber threat intelligence research that has progressively moved from artifact-centric exchange toward meaning-centric analysis. Recent surveys of intelligence mining document a maturation of the field from the extraction of isolated indicators toward the recovery of relationships, behaviors, and campaign-level structure from unstructured sources (Sun et al., 2023). The present work advances along this trajectory by treating correlation itself, rather than extraction alone, as the central analytic act, and by grounding that correlation in learned semantic representations that tolerate the paraphrase and multilingual variation endemic to coalition reporting. A distinguishing feature of the approach is its explicit coupling of distributed and symbolic representations: neither the embedding geometry nor the extracted entity graph is treated as authoritative in isolation, and the fusion of the two is what permits confident disposition of relationships that either signal alone would leave ambiguous. This dual representation also addresses a persistent tension in operational intelligence between the desire for machine-actionable structure and the reality that the richest reporting arrives as prose. By encoding narratives holistically while simultaneously recovering their discrete observables, the pipeline preserves the interpretive nuance of free text without sacrificing the precise, indexable artifacts that downstream defensive systems require, thereby serving both automated consumption and human comprehension from a single processing path rather than forcing a premature commitment to one at the expense of the other, a commitment that has historically fragmented sharing architectures.

The operational significance of the approach becomes clearest when correlation is framed not as an end in itself but as an instrument for constructing and maintaining shared situational awareness across a distributed coalition (Endsley, 1995). In such a setting, no single partner observes the entire threat landscape, and the intelligence value of the coalition arises precisely from the composition of partial views held by its members. Semantic correlation is the mechanism by which those partial views are stitched into a common operating picture, and the latency at which it operates determines whether that picture reflects the current state of adversary activity or a stale approximation of it. The design's emphasis on real-time processing, incremental clustering, and cached encoding is therefore not incidental engineering but a direct response to the temporal requirement that shared awareness be current to be actionable. At the same time, the confidence-banded output acknowledges that shared awareness is degraded as much by false certainty as by ignorance, and that a system which asserts spurious correlations erodes the trust on which coalition sharing depends. By surfacing relationships with calibrated confidence and transparent provenance, the pipeline positions the human analyst as the final arbiter while relieving that analyst of the mechanical burden of exhaustive comparison. The discussion thus reframes the technical contribution as a socio-technical one, in which machine correlation and human judgment are deliberately partitioned according to their respective strengths and neither is asked to perform the other's role.

IX. LIMITATIONS AND THREATS TO VALIDITY

Several limitations qualify the interpretation of the results and the generality of the proposed design. Foremost, the evaluation is illustrative rather than comparative, and the absence of a shared, sensitive benchmark of correlated coalition reporting means that the reported regularities describe characteristic behavior rather than establish superiority over alternative systems. The confidence bands, while operationally intuitive, rest on numeric boundaries that are contingent on the particular encoder, corpus, and threshold calibration employed, and they should be regarded as a

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

representational device to be recalibrated for any concrete deployment rather than as fixed constants. A further construct-validity concern arises from the reliance on semantic similarity as a proxy for operational relatedness; two reports may be judged similar by a language model yet refer to unrelated events, and conversely genuinely related reports may be encoded far apart when they emphasize different facets of a campaign. The fusion with symbolic extraction mitigates but does not eliminate this gap, since the extraction models themselves are imperfect and inherit domain and language biases from their training data. Internal validity is additionally threatened by the dependence of clustering outcomes on hyperparameters governing density and minimum cluster size, whose optimal values shift with stream composition. Because these parameters interact with the adaptive thresholds in the scoring stage, tuning one in isolation can produce misleading conclusions about the contribution of any single component to overall correlation quality, and disentangling their joint effect would require a far more controlled ablation than the present characterization affords.

External validity is constrained by the characteristics of coalition data that a controlled evaluation only partially reproduces. Real shared reporting is subject to classification restrictions, inconsistent formatting, adversarial manipulation, and pronounced imbalance across partners in both volume and quality, none of which is fully captured by a synthesized stream. Adversarial evasion deserves particular emphasis, since an opponent aware that semantic correlation is in use may deliberately craft reports, or the underlying activity, to defeat both embedding proximity and entity extraction, for example by fragmenting a campaign across superficially unrelated descriptions. The pipeline offers no explicit defense against such poisoning of the shared stream, and its reliance on learned representations introduces the possibility that maliciously constructed inputs could steer clustering toward false consolidation or unwarranted fragmentation. Latency measurements, moreover, were characterized under controlled arrival rates, and sustained surge conditions or degraded network links between partners could erode the real-time guarantees on which the operational argument depends. There are also privacy and sovereignty considerations that a purely technical evaluation cannot address: partners may be unwilling to submit narratives rich enough for semantic encoding, and any central index of encoded reports concentrates sensitive information that becomes an attractive target. These threats do not invalidate the approach, but they delimit the conditions under which its demonstrated behavior can be expected to hold, and they identify the assurances a production deployment would need to establish before the pipeline could be trusted in a defensive coalition operating under genuine adversarial pressure.

X. DEPLOYMENT CONSIDERATIONS FOR COALITION ENVIRONMENTS

Practical deployment of the correlation service depends first on its ability to interoperate with the structured sharing standards and platforms already entrenched across coalition partners. Threat information is routinely exchanged through standardized representation and transport formats whose schemas encode observables, relationships, and campaign metadata in machine-readable form (Barnum, 2014), and behavioral reporting is increasingly anchored to shared adversary-technique taxonomies that furnish a common vocabulary for tactics (Strom et al., 2018). The pipeline is therefore designed to consume these structured payloads directly while treating their embedded free-text narratives as the substrate for semantic encoding, so that no reporting fidelity is discarded at the boundary. Because coalition feeds arrive continuously and asynchronously, the ingestion layer is engineered as a streaming system in which normalization, encoding, and indexing proceed incrementally rather than in batch, preserving the low-latency guarantees that operational relevance demands (Tounsi & Rais, 2018). Backpressure handling, idempotent deduplication of retransmitted reports, and schema-version negotiation are treated as first-class engineering concerns, since heterogeneity across partner platforms is the norm rather than the exception (Wagner et al., 2019). Aligning correlated hypotheses with kill-chain stage models further allows the service to emit output that downstream defensive tooling can act upon without translation (Hutchins et al., 2011). This standards-aware ingestion design ensures that semantic correlation augments the existing sharing fabric (Sun et al., 2023) rather than requiring partners to abandon the investments and workflows they already maintain.

Beyond format interoperability, sustained coalition operation hinges on the disciplined management of trust, provenance, and the differential reliability of contributing sources. Reporting quality varies markedly across partners, and empirical study of shared intelligence has documented pervasive inconsistency, staleness, and duplication that a naive correlator would propagate uncritically (Tounsi & Rais, 2018; Wagner et al., 2019). The service therefore attaches immutable provenance to every ingested indicator and carries it through encoding, extraction, and consolidation, so that each synthesized hypothesis exposes not only what was concluded but which partners and which original reports substantiate it (Barnum, 2014). Source weighting is applied so that corroboration from historically reliable contributors elevates a hypothesis more than an equal count of low-confidence submissions, and this weighting interacts with the confidence bands to temper machine assertions that rest on thin or homogeneous evidence. Privacy and sovereignty constraints impose a further discipline, because partners are frequently prohibited from exposing raw narratives outside their jurisdiction (Rahman et al., 2020). The architecture accommodates this by supporting correlation over encoded representations that need not reveal underlying text (Cer et al., 2018; Reimers & Gurevych, 2019), and by permitting partners to withhold or redact fields while still contributing to shared clustering. Such controls preserve the information-theoretic value of aggregation (Shannon, 1948) without compelling any participant to surrender data-handling authority, which is the precondition for continued voluntary participation in a coalition sharing arrangement operating under divergent national policy regimes.

Finally, a correlation service embedded in national defensive infrastructure requires continuous monitoring, human feedback, and explicit operational governance to remain trustworthy as its environment evolves. Learned encoders are trained on distributions that drift as adversary tradecraft, reporting conventions, and language usage change, and undetected representational drift silently degrades correlation quality (Sarker et al., 2020). The deployment therefore instruments the embedding space and the clustering statistics, tracking indicators such as neighborhood density and band occupancy so that drift is surfaced and periodic re-encoding or fine-tuning can be triggered before analytic reliability erodes (Devlin et al., 2019; Gao et al., 2021). Analyst dispositions of proposed correlations are captured as labeled feedback that refines source weights, recalibrates band thresholds, and supplies scarce supervision for the extraction models (Bridges et al., 2013; Husari et al., 2017), closing a loop in which human judgment continually corrects machine inference rather than merely consuming it. This feedback discipline respects the cognitive limits that motivated consolidation in the first place (Miller, 1956) and reinforces the shared situational awareness the coalition seeks (Endsley, 1995). Governance overlays these mechanisms with accountability: versioned models, auditable decision trails, and defined escalation paths ensure

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

that when the service asserts or omits a correlation, the reasoning can be reconstructed and challenged. Treating the correlation service as a governed, monitored, and revisable operational system (Manning & Raghavan, 2008; Sun et al., 2023), rather than a static analytic artifact, is what allows it to be depended upon across the long horizons of coalition defense.

XI. CONCLUSION AND FUTURE WORK

This paper has argued that the correlation of shared cyber threat indicators in national defense is fundamentally a problem of meaning rather than of matching, and it has developed a real-time pipeline that treats semantic understanding as the organizing principle of the analytic workflow. The pipeline composes interchangeable stages for semantic encoding, similarity scoring, symbolic entity and relationship extraction, and density-based clustering and consolidation, fusing distributed and symbolic representations so that correlation decisions rest on converging evidence rather than on any single fragile signal. An illustrative evaluation demonstrated that the system links semantically equivalent reports lacking lexical overlap, contracts high-redundancy streams into a bounded set of corroborated hypotheses, and stratifies its output into interpretable confidence bands aligned with distinct operational meanings. The interpretation connected these behaviors to established theory, showing that consolidation reduces the entropy of the analyst's input, that correlation advances the perception, comprehension, and projection components of situation awareness, and that volume contraction is a precondition for respecting the finite capacity of human information processing. Taken together, these contributions position semantic correlation not as an incremental improvement to indicator exchange but as a reframing of the analyst's relationship to a high-velocity, distributed reporting environment, in which machine understanding and human judgment are partitioned according to their respective strengths and the coalition's shared operating picture is assembled continuously from partial views that no single partner could resolve alone.

Future work should proceed along several complementary directions. The most pressing is the construction of shared, appropriately sanitized benchmarks of correlated coalition reporting, without which comparative evaluation and principled threshold calibration will remain out of reach. A second direction concerns robustness to adversarial manipulation, encompassing both the detection of poisoned submissions and the design of encoders and extractors whose correlation decisions degrade gracefully under deliberate evasion. A third avenue is the extension of the pipeline from pairwise and cluster-level correlation toward explicit temporal reasoning about campaign evolution, so that the system projects likely adversary progression rather than only consolidating past observations. Advances in domain-adapted and contrastively trained representations promise to narrow the gap between semantic similarity and operational relatedness, and their integration warrants systematic study under the sensitivity constraints of the defense setting. Privacy-preserving architectures, including federated encoding and correlation over representations that never expose raw narratives, merit particular attention given the sovereignty concerns that govern coalition sharing. Finally, closer study of the human-machine partition, through operational trials that measure how confidence-banded correlation affects analyst decision quality and workload, would ground the cognitive claims advanced here in empirical practice. Pursued together, these directions would move semantic correlation from a demonstrated capability toward a dependable component of national defensive infrastructure, one capable of sustaining shared understanding at the tempo and scale that contemporary threat environments demand of the coalitions that confront them.

REFERENCES

1) Adebayo, A., Adepoju, P. A., & Ozowara, D. E. (2023). Conceptual model for cyber risk pricing and insurance structuring in health technology platforms. Gyanshauryam, International Scientific Refereed Research Journal, 6(6). https://doi.org/10.32628/GISRRJ
2) Adebayo, A., Anunagba, C. O., & Ozowara, D. E. (2025). A review of zero trust security models and cost effectiveness in healthcare infrastructure. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2269–2283. https://doi.org/10.62225/2583049X.2025.5.6.6051
3) Adeyelu, O. O. (2023). A Risk-Tiered Oversight Model for Unmanned Aircraft Operations in Shared Controlled Airspace Over Nigerian Airports. International Journal of Scientific Research in Civil Engineering, 7(4), 147–184. https://doi.org/10.32628/IJSRCE237419
4) Agbabiaka, J., Okonkwo, C. S., Ogunwole, O., Mayo, W., & Okeke, O. T. (2019). Supply Chain Risk Management Model for EPC and Gas Processing Projects. IRE Journals, 3(2), 968–980. https://doi.org/10.64388/IREV3I2-1713124
5) Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In Mining text data (pp. 77–128). Springer.
6) Akanbi, O., Ganiu, O. S., & Sunday, E. A. (2025). A multi-agent AI framework for swarm intelligence in autonomous mobile robot (AMR) fleet coordination. International Journal of Scientific Research in Humanities and Social Sciences, 2(1), 81–109.
7) Akeju, B., Edivri, J., Ogbole, J. I., Okoruwa, P. O., Fadayomi, O., & Abolaji, T. O. (2018). Conceptual model for insider threat classification and risk modeling in complex digital systems. Iconic Research and Engineering Journals, 1(9), 476–492. https://doi.org/10.64388/IREV1I9-1713778
8) Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2023a). Advances in Predictive Analytics Models for Student Retention and Institutional Risk Management Systems. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2692–2711. https://doi.org/10.62225/2583049X.2023.3.6.5990
9) Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2023b). A review of API governance and risk prioritization frameworks in modern financial institutions. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(10), 395–433. https://doi.org/10.32628/CSEIT2361072
10) Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2023c). A conceptual framework for continuous cloud misconfiguration monitoring and enterprise risk mitigation strategies. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(10), 373–394. https://doi.org/10.32628/CSEIT2361071
11) Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2024a). A conceptual framework for enterprise data sensitivity classification and regulatory traceability mechanisms. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 3103–3124. https://doi.org/10.62225/2583049X.2024.4.6.5991

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

12) Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2024b). Advances in HIPAA compliant data architecture and secure analytics frameworks for community healthcare organizations. Shodhshauryam, International Scientific Refereed Research Journal, 7(2), 277–324. https://doi.org/10.32628/SHISRRJ2472163
13) Aliliele, C., Mbonu, I. S., Uzoka, E., & Iwuanyanwu, U. (2025a). A review of AI assisted continuous auditing systems in technology risk and cybersecurity oversight. Gyanshauryam, International Scientific Refereed Research Journal, 8(4), 210–250. https://doi.org/10.32628/GISRRJ258369
14) Aliliele, C., Mbonu, I. S., Uzoka, E., & Iwuanyanwu, U. (2025b). Advances in data lakehouse governance architectures for enterprise data loss prevention and compliance assurance. Shodhshauryam, International Scientific Refereed Research Journal, 8(4), 171–213. https://doi.org/10.32628/SHISRRJ258474
15) Amadi, C. (2025). Accelerating national defense: Using large language models (LLM) and NLP for real-time semantic correlation and de-duplication of shared threat indicators. SSRN. https://ssrn.com/abstract=5940895
16) Aminu-Ibrahim, A. Y., & Ogbete, J. C. (2023). Healthcare Infrastructure as a Public Health Intervention Using Evidence from Large Laboratory Networks. Shodhshauryam, International Scientific Refereed Research Journal, 6(1), 256–286. https://doi.org/10.32628/SHISRRJ23678
17) Aminu-Ibrahim, A., Ogbete, J. C., & Iwuanyanwu, O. C. (2025a). Sustainable Healthcare Infrastructure Performance Metrics for Long-Term Asset Management and Value Creation. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(4), 566–601. https://doi.org/10.32628/CSEIT251116277
18) Aminu-Ibrahim, A. Y., Ogbete, J. C., & Iwuanyanwu, O. C. (2025b). Infrastructure Resilience Planning for National Diagnostic Systems Under Public Health Stress Conditions. Gyanshauryam, International Scientific Refereed Research Journal, 8(1), 340–381. https://doi.org/10.32628/GISRRJ2582311
19) Annan, A. O. (2022). Data privacy governance models for cross-border digital platforms. International Journal of Multidisciplinary Research and Growth Evaluation, 3(6), 949–964.
20) Annan, A. O. (2024). Algorithmic accountability and trade secret protection in artificial intelligence. Shodhshauryam, International Scientific Refereed Research Journal, 7(5), 315–347.
21) Annan, A. O. (2025a). Automated decision-making and anti-discrimination compliance under U.S. law. International Journal of Advanced Multidisciplinary Research and Studies, 5(2), 2522–2540.
22) Annan, A. O. (2025b). Cybersecurity compliance as a source of competitive advantage in technology markets. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(5), 456–487.
23) Arumosoye, O. M., & Obriki, O. D. (2019). Systematic Review of Near-Miss and Hazard Observation Data Utilization in Industrial Safety Management. IRE Journals, 3(2), 981–999. https://doi.org/10.64388/IREV3I2-1714417
24) Arumosoye, O. M., & Obriki, O. D. (2021). Organizational Learning-Based Conceptual Maturity Model for Continuous Safety Performance Improvement. International Journal of Multidisciplinary Research and Growth Evaluation, 2(1), 970–980. https://doi.org/10.54660/.IJMRGE.2021.2.1.970-980
25) Arumosoye, O. M., & Obriki, O. D. (2023). Conceptual Model for Emergency Response Readiness and Capability in Energy and Process Facilities. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(3), 897–917. https://doi.org/10.32628/CSEIT25112791
26) Arumosoye, O. M., Obriki, O. D., & Ozobu, C. O. (2026). Systematic Review of Predictive Safety Analytics Applications in LNG Projects with ESG Implications. Global Journal of Engineering and Technology Review, 2(2), 61–73. https://doi.org/10.65150/EP-gjetr/V2E2/2026-05
27) Asiedu, C. S. (2023). Diagnostic integration in infectious disease management: A review of microbiology, virology, and serology workflows. International Journal of Medical Evaluation and Physical Report, 7(4), 173–195. https://doi.org/10.56201/ijmepr.v7.no4.2023.pg173.195
28) Asiedu, W., & Quainoo, R. (2023a). How far can energy harvesting take us? A systematic review of radio frequency strategies for energy autonomous sensing. Shodhshauryam, International Scientific Refereed Research Journal, 6(1), 448–466.
29) Asiedu, W., & Quainoo, R. (2023b). Toward maintenance free wireless infrastructure: Simulating performance and reliability in large scale intermittently powered IoT networks. Gyanshauryam, International Scientific Refereed Research Journal, 6(1), 489–510.
30) Asiedu, W., & Quainoo, R. (2024). Rethinking energy, reliability, and latency trade-offs in green communication for next generation IoT. International Journal of Multidisciplinary Research and Growth Evaluation, 5(6), 1987–1994.
31) Asiedu, W., & Quainoo, R. (2025). GleanCast: Epoch-aligned reliable broadcast in batteryless intermittent sensor networks. International Journal of Engineering and Modern Technology, 11(12), 205–218. https://doi.org/10.56201/ijemt.vol.11.no12.2025.pg205.218
32) Asiedu, W., Quainoo, R., & Asiedu, A. (2025). A conceptual framework for characterizing fundamental energy, reliability, and latency trade-offs in green communication paradigms for next-generation IoT. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2447–2453. https://doi.org/10.62225/2583049X.2025.5.6.6479
33) Asiedu, W., Quainoo, R., & Asiedu, A. (2026). Predictive power management for intermittent IoT devices using lightweight machine learning under uncertain energy harvest. Gulf Journal of Engineering and Technology, 2(4), 108–118.
34) Atta, S., Tofah, P., Kankam, M. A., & Adenuga, O. M. (2020). A critical review of NDT-based integrity management and corrosion risk assessment strategies for refinery process equipment. International Journal of Engineering and Modern Technology, 6(2), 19–46. https://doi.org/10.56201/ijemt.vol.6.no2.2020.pg19.46

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

35) Atta, S., Kerubo, M. L., Sabisa, N. E., Adu-Gyamfi, D., Appiah, S., & Onyishi, E. (2025). Environmental corrosion and long-term degradation of crystalline silicon solar cells: Mechanisms, climate effects and mitigation strategies. Current Journal of Applied Science and Technology, 44(10), 9–18.
36) Badmus, O., Dosunmu, A. A., & Ozowara, D. E. (2018). A systematic review of CI/CD pipeline strategies in Salesforce DevOps: Tools, practices, and deployment outcomes. IRE Journals, 2(6).
37) Badmus, O., Dosunmu, A. A., & Ozowara, D. E. (2019a). A governance framework for Salesforce platform management in regulated healthcare environments. IRE Journals, 3(6).
38) Badmus, O., Dosunmu, A. A., & Ozowara, D. E. (2019b). A conceptual model for ETL design and data integration in Salesforce-centric enterprise architectures. IRE Journals, 3(5).
39) Badmus, O., Jooda, D., Ozowara, D. E., & Anunagba, C. O. (2022a). A framework for Git-based version control and code review governance in Salesforce development teams. Shodhshauryam, International Scientific Refereed Research Journal, 5(2), 473–493.
40) Badmus, O., Jooda, D., Ozowara, D. E., & Anunagba, C. O. (2022b). A systematic review of MuleSoft ESB integration patterns in multi-cloud Salesforce architectures. Gyanshauryam, International Scientific Refereed Research Journal, 5(3), 462–481.
41) Badmus, O., Jooda, D., & Anunagba, C. O. (2024). A privacy-by-design framework for role-based security architecture in Salesforce environments handling sensitive personal data. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 3244–3254. https://doi.org/10.62225/2583049X.2024.4.6.6243
42) Badmus, O., Dosunmu, A. A., & Anunagba, C. O. (2025). A governance framework for AI-assisted CRM workflows: Agentforce deployment, risk management, and organizational readiness in enterprise Salesforce environments. International Journal of Scientific Research in Humanities and Social Sciences, 2(1), 59–80. https://doi.org/10.32628/IJSRHSS
43) Barnum, S. (2014). Standardizing cyber threat intelligence information with the Structured Threat Information eXpression (STIX). MITRE Corporation.
44) Bello, A. D., Elebe, O., Hammed, N. I., Omoegun, G. O., & Fadayomi, O. (2024). A cybersecurity risk management and regulatory compliance framework for financial institutions. Iconic Research and Engineering Journals. https://www.irejournals.com/paper-details/1713553
45) Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
46) Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146.
47) Borah, S., Aliliele, K. C., Rakshit, S., & Vajjhala, N. R. (2022). Applications of artificial intelligence in software testing. In P. K. Mallick et al. (Eds.), Cognitive informatics and soft computing (Lecture Notes in Networks and Systems, Vol. 375, pp. 727–736). Springer. https://doi.org/10.1007/978-981-16-8763-1_60
48) Bridges, R. A., Jones, C. L., Iannacone, M. D., Testa, K. M., & Goodall, J. R. (2013). Automatic labeling for entity extraction in cyber security. arXiv:1308.4941.
49) Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
50) Cer, D., Yang, Y., Kong, S.-Y., Hua, N., Limtiaco, N., St. John, R., ... Kurzweil, R. (2018). Universal sentence encoder. arXiv:1803.11175.
51) Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. Proceedings of EMNLP, 670–680.
52) Dagodzo, D. (2018a). A Conceptual Framework for UAV Integration into National Power Grid Inspection Programs. IRE Journals, 2(5), 391–412. https://doi.org/10.64388/IREV2I5-1716082
53) Dagodzo, D. (2018b). A Review of UAV Applications in Electrical Transmission Line Inspection: Methods, Technologies, and Challenges. IRE Journals, 2(6), 234–254. https://doi.org/10.64388/IREV2I6-1716083
54) Dagodzo, D., & Ahiaeke Patrick, M. C. (2021). An Integrated Framework for UAV, LiDAR, and GIS in Infrastructure Corridor Management. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(5), 497–524. https://doi.org/10.32628/CSEIT217566
55) Dagodzo, D., & Ahiaeke Patrick, M. C. (2022). A Review of Right-of-Way Encroachment Detection Methods Using Geospatial Technologies. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(1), 638–667. https://doi.org/10.32628/CSEIT2281226
56) Dagodzo, D., Patrick, M. C. A., & Aliliele, C. (2022). AI and deep learning for vegetation classification in power corridor management: A review. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(1), 668–698. https://doi.org/10.32628/CSEIT2281227
57) Dagodzo, D., & Ahiaeke Patrick, M. C. (2025). A Framework for National-Scale UAV Deployment in Power Infrastructure: Lessons from Developing Economies. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(4), 779–824. https://doi.org/10.32628/CSEIT251116286
58) Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186. https://doi.org/10.18653/v1/N19-1423
59) Dosunmu, A. A., & Ogundele, P. O. (2019). Security audit and enterprise risk assessment frameworks for resilient information systems. IRE Journals, 3(5).
60) Dosunmu, A. A., & Ogundele, P. O. (2021). Incident response and digital forensics strategies for rapid cyber attack containment. Gyanshauryam, International Scientific Refereed Research Journal, 4(4), 239–258.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

61) Dosunmu, A. A., & Ogundele, P. O. (2022). Threat intelligence integration frameworks supporting proactive enterprise cybersecurity decision making. Gyanshauryam, International Scientific Refereed Research Journal, 5(3), 397–416.
62) Dosunmu, A. A., & Ogundele, P. O. (2023). Cyber threat actor analysis models for strategic enterprise security planning. Shodhshauryam, International Scientific Refereed Research Journal, 6(5), 513–531. https://doi.org/10.32628/SHISRRJ
63) Dosunmu, A. A., & Ogundele, P. O. (2024a). Threat-informed defence engineering models for measuring security control effectiveness at scale. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2847–2858.
64) Dosunmu, A. A., & Ogundele, P. O. (2024b). Breach and attack simulation frameworks for continuous validation of enterprise security controls. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(3), 1100–1119. https://doi.org/10.32628/IJSRCSEIT
65) Dosunmu, A. A., & Ogundele, P. O. (2024c). Cyber risk quantification models for prioritizing enterprise security investment decisions. International Journal of Multidisciplinary Research and Growth Evaluation, 5(6), 1777–1785.
66) Dosunmu, A. A., & Ogundele, P. O. (2024d). Enterprise scale continuous security validation models for regulated digital infrastructures. International Journal of Scientific Research in Humanities and Social Sciences, 1(2), 929–945. https://doi.org/10.32628/IJSRSSH
67) Dosunmu, A. A., & Ogundele, P. O. (2025a). Security orchestration and automation models for accelerating incident detection and response. Computer Science and IT Research Journal, 6(11), 878–894. https://doi.org/10.51594/csitrj.v6i11.2163
68) Dosunmu, A. A., & Ogundele, P. O. (2025b). Adversary simulation design frameworks for proactive cyber defense in complex environments. International Journal of Computer Science and Mathematical Theory, 11(12), 193–209. https://doi.org/10.56201/ijcsmt.vol.11.no12.2025.pg193.209
69) Dosunmu, A. A., & Ogundele, P. O. (2025c). Cyber defense performance measurement frameworks for executive and board-level security governance. Computer Science and IT Research Journal, 6(11), 895–913.
70) Eboh, E. E., & Aliliele, C. (2023). Portfolio Risk Optimization Models for Venture Capital Investment Strategy and Fraud Exposure Analysis. Iconic Research and Engineering Journals, 7(4), 759–792. https://doi.org/10.64388/IREV7I4-1715305
71) Eboh, E. E., & Aliliele, C. (2024). AI-Driven Data Analytics Framework for Risk Assessment and Detection of Venture Capital and Private Equity Investment Fraud in U. S. Capital Markets. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(6), 2624–2665. https://doi.org/10.32628/CSEIT2410787
72) Eboh, E. E., & Aliliele, C. (2025a). Predictive Analytics Systems for Investment Risk Monitoring in SME FinTech Companies and Cross-Border Capital Markets. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(2), 3969–4010. https://doi.org/10.32628/CSEIT25113398
73) Eboh, E. E., & Aliliele, C. (2025b). Advanced Financial Due Diligence Models for Preventing Venture Capital Investment Fraud and Strengthening Investor Protection. Shodhshauryam, International Scientific Refereed Research Journal, 8(5), 153–197. https://doi.org/10.32628/SHISRRJ258526
74) Edivri, J., & Oteri, O. (2022). Predictive capacity planning and resource utilization forecasting models for multi-program and multi-stakeholder IT portfolios. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(4), 826–845.
75) Edivri, J., Ogbole, J. I., Okoruwa, P. O., Fadayomi, O., Abolaji, T. O., & Akeju, B. (2026). Conceptual model for integrated human and machine identity governance in cloud-based security architectures [Manuscript; venue not yet indexed].
76) Ekechi, N. V., Anunagba, C. O., & Ozowara, D. E. (2022). Advances in financial forensics techniques for detecting cyber-enabled fraud in telemedicine services. Shodhshauryam, International Scientific Refereed Research Journal, 5(1), 428–450. https://doi.org/10.32628/SHISRRJ
77) Ekechi, N. V., Ozowara, D. E., & Anunagba, C. O. (2026). Conceptual framework for AI governance, data privacy compliance, and financial sustainability in digital health. Computer Science and IT Research Journal, 7(4), 275–299.
78) Elebe, O., Hammed, N. I., Omoegun, G. O., Fadayomi, O., & Bello, A. D. (2023). An advanced machine learning model for detecting synthetic identity fraud in e-commerce platforms. ResearchGate. https://www.researchgate.net/profile/Adepeju-Bello-2
79) Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64.
80) Essandoh, S., Sakyi, J. K., Ibrahim, A. K., Okafor, C. M., & Wedraogo, L. (2025). Artificial intelligence and the future of work: Impacts on employment and job roles. International Journal of Multidisciplinary Futuristic Development, 6(1), 31–41. https://doi.org/10.54660/IJMFD.2025.6.1.31-41
81) Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of KDD, 226–231.
82) Eyetsemitan, R. A., Ambali, K. B., Oyeleye, A. O., & Fadayomi, O. (2020). Multi-Stakeholder Governance Alignment in Joint Venture Operations: A Conceptual Framework for Coordinating Business Processes in Highly Regulated Environments. IRE Journals, 4(4), 418–441. https://doi.org/10.64388/IREV4I4-1716955
83) Eyetsemitan, R. A., Oyeleye, A. O., Ambali, K. B., & Fadayomi, O. (2024). Data-Driven Process Optimization in Micro-Enterprises: A Conceptual Framework for Funnel Analysis and Bottleneck Identification. International Journal of Multidisciplinary Research and Growth Evaluation, 5(6), 1950–1968. https://doi.org/10.54660/IJMRGE.2024.5.6.1950-1968
84) Fadayomi, O., Bello, A. D., Elebe, O., Hammed, N. I., & Omoegun, G. O. (2021). An integrated cybersecurity and anti-money laundering governance framework for financial crime prevention. ResearchGate. https://www.researchgate.net/profile/Adepeju-Bello-2
85) Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. Proceedings of EMNLP, 6894–6910.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

86) Ghazi, Y., Anwar, Z., Mumtaz, R., Saleem, S., & Tahir, A. (2018). A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources. Proceedings of the International Conference on Frontiers of Information Technology, 129–134.
87) Husari, G., Al-Shaer, E., Ahmed, M., Chu, B., & Niu, X. (2017). TTPDrill: Automatic and accurate extraction of threat actions from unstructured text of CTI sources. Proceedings of ACSAC, 103–115.
88) Hutchins, E. M., Cloppert, M. J., & Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. Leading Issues in Information Warfare & Security Research, 1(1), 80–106.
89) Jooda, D., & Smart, B. D. (2024a). An ethical and secure generative AI deployment model for construction and infrastructure systems. Shodhshauryam, International Scientific Refereed Research Journal, 7(2), 349–359. https://doi.org/10.32628/SHISRRJ
90) Jooda, D., & Smart, B. D. (2024b). Risk-adaptive cybersecurity compliance automation model (RACCA): Continuous assessment, dynamic threshold management, and integrated remediation for enterprise security programs. Shodhshauryam, International Scientific Refereed Research Journal, 7(2), 360–374. https://doi.org/10.32628/SHISRRJ
91) Joshi, A., Lal, R., Finin, T., & Joshi, A. (2013). Extracting cybersecurity related linked data from text. Proceedings of the IEEE International Conference on Semantic Computing, 252–259.
92) Komi, N. M., & Adeniji, I. O. (2020). Co-Optimizing Technical Performance and Community Affordability in Smart Solar Microgrids for Underserved Regions. International Journal of Engineering and Modern Technology, 6(3), 173–212. https://doi.org/10.56201/ijemt.vol.6.no3.2020.pg173.212
93) Komi, N. M. (2021). Hybrid Deep Learning for Wind Power Forecasting: From Grid Stability to Market Equity for Small Generators. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 989–1014. https://doi.org/10.54660/.IJMRGE.2021.2.6.989-1014
94) Komi, N. M., & Adamolekun, A. (2021). Interpretable Machine Learning for Early Failure Prediction in Distributed Renewable Energy Assets. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 1015–1038. https://doi.org/10.54660/.IJMRGE.2021.2.6.1015-1038
95) Komi, N. M., & Adamolekun, A. (2022). Hierarchical, Decentralized, or Hybrid? A Systematic Review of Control Architectures for Distributed Energy Resources. International Journal of Engineering and Modern Technology, 8(5), 141–195. https://doi.org/10.56201/ijemt.v8.no5.2022.pg141.195
96) Komi, N. M., & Ganiu, O. S. (2022). Grid-Connected Battery Storage in Hot and Weak-Grid Regions: Advances, Degradation Challenges, and Lifecycle Equity. International Journal of Multidisciplinary Research and Growth Evaluation, 3(6), 1072–1095. https://doi.org/10.54660/.IJMRGE.2022.3.6.1072-1095
97) Komi, N. M. (2023). When Accurate Forecasts Are Not Fair Forecasts: Deep Learning Load Prediction and Equity Across Districts. Shodhshauryam, International Scientific Refereed Research Journal, 6(6), 476–546. https://doi.org/10.32628/SHISRRJ247102
98) Komi, N. M., & Adeniji, I. O. (2023). A Tiered Digital Twin that Brings Predictive Diagnostics to Resource-Constrained Renewable Energy Sites. International Journal of Engineering and Modern Technology, 9(3), 287–347. https://doi.org/10.56201/ijemt.v9.no3.2023.pg287.347
99) Komi, N. M., & Ganiu, O. S. (2023). Edge Intelligence for Resilient Microgrid Control: Advances, Energy Sovereignty, and Open Challenges. Gyanshauryam, International Scientific Refereed Research Journal, 6(3), 525–586. https://doi.org/10.32628/GISRRJ236339
100) Kumuyi, O., Akeju, B., Uzoka, E., & Ozowara, D. E. (2023). Framework for blockchain-based cross-border data exchange and regulatory transparency. International Journal of Multidisciplinary Futuristic Development, 4(1), 99–113.
101) Kumuyi, O., Uzoka, E., Akeju, B., & Ozowara, D. E. (2024). Framework for privacy-focused digital identity verification supporting financial inclusion in Africa. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 3150–3162. https://doi.org/10.62225/2583049X.2024.4.6.6005
102) Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2019). Implementation of Active Directory for efficient management of enterprise networks. IRE Journals, 3(4). https://doi.org/10.64388/IREV3I4-1717206
103) Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2022a). Human-in-the-loop machine learning: A state of the art. Journal of Frontiers in Multidisciplinary Research, 3(1), 656–669.
104) Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2022b). Navigating digital transformation: Best practices for cloud migration strategies in the enterprise. Journal of Frontiers in Multidisciplinary Research, 3(1), 643–655.
105) Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2023). A comprehensive survey on ServiceNow for IT service management. International Journal of Multidisciplinary Research and Growth Evaluation, 4(6), 1532–1545.
106) Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2024). Keeping humans in the loop: Human-centered automated annotation with generative AI. International Journal of Multidisciplinary Futuristic Development, 5(1), 81–95.
107) Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2025). Migration of applications and information systems to cloud computing infrastructure: Lessons from a South African retail bank. International Journal of Multidisciplinary Research and Growth Evaluation, 6(6), 1361–1375.
108) Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2026). On the disagreement problem in human-in-the-loop federated machine learning. International Journal of Multidisciplinary Research and Growth Evaluation, 7(3), 178–192.
109) Liao, X., Yuan, K., Wang, X., Li, Z., Xing, L., & Beyah, R. (2016). Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence. Proceedings of the ACM CCS, 755–766.
110) Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Corresponding Author: Chukwunenye Amadi

111) Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.
112) Mayo, W., Ogbole, J. I., Okoruwa, P. O., & Babatope, O. M. (2021). Designing an AI-predictive maintenance model for e-commerce systems using machine learning and cloud analytics. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(5), 416–440. https://doi.org/10.32628/IJSRCSEIT
113) Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Oluoha, O. M. (2018). A conceptual framework for legal and ethical risk modeling in enterprise data protection governance systems. Iconic Research and Engineering Journals, 2(2), 207–226. https://doi.org/10.64388/IREV2I2-1714911
114) Mbonu, I. S., Iwuanyanwu, U., Uzoka, E., & Oluoha, O. M. (2019a). Advances in enterprise log analytics and automated incident response architectures using Python and SIEM platforms. Iconic Research and Engineering Journals, 3(2), 1000–1019. https://doi.org/10.64388/IREV3I2-1714915
115) Mbonu, I. S., Aliliele, C., Uzoka, E., & Oluoha, O. M. (2019b). A review of comparative data protection regulations and secure cloud implementation strategies across jurisdictions. Iconic Research and Engineering Journals, 2(9), 482–501. https://doi.org/10.64388/IREV2I9-1714912
116) Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2020a). A conceptual framework for agile supply chain digital transformation with embedded IT risk and ISO compliance controls. Iconic Research and Engineering Journals, 3(11), 566–593. https://doi.org/10.64388/IREV3I11-1714916
117) Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Uzoka, E. (2020b). A review of identity and access management integration strategies in hybrid and multi cloud environments. International Journal of Multidisciplinary Research and Growth Evaluation, 1(5), 795–810. https://doi.org/10.54660/.IJMRGE.2020.1.5.795-810
118) Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2020c). Advances in infrastructure as code governance for secure Terraform based enterprise cloud deployments. International Journal of Multidisciplinary Research and Growth Evaluation, 1(5), 811–828. https://doi.org/10.54660/.IJMRGE.2020.1.5.811-828
119) Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2021a). A review of VoIP forensic analytics models for financial fraud detection and regulatory compliance monitoring. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 711–730. https://doi.org/10.54660/.IJMRGE.2021.2.6.711-730
120) Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Uzoka, E. (2021b). Advances in artificial intelligence techniques for secure software testing and automated regression control mechanisms. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(5), 468–496. https://doi.org/10.32628/CSEIT217565
121) Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Uzoka, E. (2021c). A conceptual framework for risk based business intelligence architecture in financial technology platforms. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 731–746. https://doi.org/10.54660/.IJMRGE.2021.2.6.731-746
122) Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2022a). A review of data protection impact assessment models in multi cloud financial infrastructure systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(1), 589–623. https://doi.org/10.32628/CSEIT25442
123) Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Uzoka, E. (2022b). A conceptual framework for AI enabled IT general controls and SOX audit automation processes. Gyanshauryam, International Scientific Refereed Research Journal, 5(5), 384–414. https://doi.org/10.32628/GISRRJ2256239
124) Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2022c). Advances in cloud identity and access governance optimization in large scale AWS enterprise environments. Shodhshauryam, International Scientific Refereed Research Journal, 5(3), 403–438. https://doi.org/10.32628/SHISRRJ225490
125) Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv:1301.3781.
126) Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111–3119.
127) Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81–97.
128) Mittal, S., Das, P. K., Mulwad, V., Joshi, A., & Finin, T. (2016). CyberTwitter: Using Twitter to generate alerts for cybersecurity threats and vulnerabilities. Proceedings of IEEE/ACM ASONAM, 860–867.
129) Obogo, S. F., Uduokhai, D. O., & Ozobu, C. O. (2019). Systematic Review of Hazard Identification and Risk Control Practices in Urban Construction Projects. Iconic Research and Engineering Journals, 2(12), 666–681. https://doi.org/10.64388/IREV2I12-1715496
130) Obogo, S. F., Arumosoye, O. M., & Obriki, O. D. (2021). Advances in Proactive Hazard Recognition and Near Miss Reporting Systems. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 835–846. https://doi.org/10.54660/IJMRGE.2021.2.6.835-846
131) Obogo, S. F., Ozobu, C. O., Nwafor, M. I., & Adio, S. A. (2025). Advances in Hazard Identification Systems for Large Scale Construction Operations. International Journal of Scientific Research in Civil Engineering, 9(6), 37–70. https://doi.org/10.32628/IJSRCE2154969
132) Obogo, S. F., Ozobu, C. O., Garba, B. M. P., & Adio, S. A. (2026). Predictive Safety Analytics Model for Early Detection of High-Risk Construction Activities. Global Journal of Engineering and Technology Review, 2(3), 92–109. https://doi.org/10.65150/EP-gjetr/V2E3/2026-03
133) Obriki, O. D., & Arumosoye, O. M. (2018). Conceptual Modeling of Data-Driven Occupational Safety Risk Control in Large-Scale Energy Infrastructure Projects. IRE Journals, 1(7), 169–189. https://doi.org/10.64388/IREV1I7-1714414

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

134) Obriki, O. D., & Arumosoye, O. M. (2019). A Conceptual Framework Linking Management Safety Walkthrough Frequency and Coverage to Safety Culture Outcomes in Mega Projects. IRE Journals, 2(8), 355–374. https://doi.org/10.64388/IREV2I8-1714416
135) Obriki, O. D., & Arumosoye, O. M. (2022). Conceptual Framework Explaining Recurrence Mechanisms of Unsafe Behaviors in High-Hazard Worksites. Shodhshauryam, International Scientific Refereed Research Journal, 5(3), 380–402. https://doi.org/10.32628/SHISRRJ225489
136) Obriki, O. D., & Arumosoye, O. M. (2023). Conceptual Framework for Proactive Hazard Identification Using Digital Safety Data Streams. Gyanshauryam, International Scientific Refereed Research Journal, 6(3), 457–481. https://doi.org/10.32628/GISRRJ236336
137) Obriki, O. D., Arumosoye, O. M., & Obogo, S. F. (2023). Advances in Continuous Hazard Monitoring Systems for Workplace Safety. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2742–2759. https://doi.org/10.62225/2583049X.2023.3.6.6075
138) Odejobi, O. D., Okonkwo, C. S., Ahiaeke Patrick, M. C., Okeke, O. T., & Mayo, W. (2025). AI-augmented secure software engineering: Leveraging deep learning for autonomous threat detection and mitigation. International Journal of Engineering and Modern Technology, 11(12), 101–121. https://doi.org/10.56201/ijemt.vol.11.no12.2025.pg101.121
139) Ogbete, J. C., Aminu-Ibrahim, A. Y., & Ambali, K. B. (2023). Lifecycle Performance Evaluation of Purpose Built Diagnostic Laboratories Supporting Long Term Healthcare Delivery. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2605–2621. https://doi.org/10.62225/2583049X.2023.3.6.5698
140) Ogbete, J. C., & Aminu-Ibrahim, A. (2024). Translating Healthcare Infrastructure Investment into Measurable Population Health and Diagnostic Outcomes. International Journal of Scientific Research in Humanities and Social Sciences, 1(2), 955–985. https://doi.org/10.32628/IJSRSSH242775
141) Ogbole, J. I., Okoruwa, P. O., Fadayomi, O., Abolaji, T. O., Edivri, J., & Akeju, B. (2021a). Conceptual model for identity-centric zero trust architecture in enterprise security governance. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(5), 393–415. https://doi.org/10.32628/IJSRCSEIT217562
142) Ogbole, J. I., Okoruwa, P. O., Babatope, O. M., & Oyewole, T. (2021b). Developing an integrated data visualization model for continuous business performance monitoring and optimization. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(5), 441–467. https://doi.org/10.32628/IJSRCSEIT
143) Ogbole, J. I., Okoruwa, P. O., Fadayomi, O., Akeju, B., Edivri, J., & Abolaji, T. O. (2025). Security analytics and digital forensics for enterprise risk management: Advances and practical implications. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2017–2028.
144) Ogunwole, O., Okonkwo, C. S., Agbabiaka, J., Mayo, W., & Okeke, O. T. (2021). Supply Chain Resilience Framework for Critical Infrastructure and Gas Processing Plants. Shodhshauryam, International Scientific Refereed Research Journal, 4(4), 444–461. https://doi.org/10.32628/SHISRRJ214462
145) Okonkwo, C. S., Ahiaeke Patrick, M. C., Okeke, O. T., & Mayo, W. (2023). Framework for Integrating IT Systems Engineering with Supply Chain Operations. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2580–2589. https://doi.org/10.62225/2583049X.2023.3.6.5500
146) Okonkwo, C. S., Agbabiaka, J., Mayo, W., & Okeke, O. T. (2024a). Framework for Secure and Scalable Supply Chain Systems Supporting National Energy Reliability. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2816–2826. https://doi.org/10.62225/2583049X.2024.4.6.5494
147) Okonkwo, C. S., Agbabiaka, J., Mayo, W., & Okeke, O. T. (2024b). Review of Digital Supply Chain Models for Cost Control and Operational Continuity. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2836–2846. https://doi.org/10.62225/2583049X.2024.4.6.5496
148) Okonkwo, C. S., Agbabiaka, J., Okeke, O. T., & Mayo, W. (2025). Framework for National-Scale Supply Chain Optimization Through Integrated IT and Procurement Systems. Gulf Journal of Advance Business Research, 3(12), 1610–1625. https://doi.org/10.51594/gjabr.v3i12.189
149) Okoruwa, P. O., Fadayomi, O., Akeju, B., Edivri, J., Ogbole, J. I., & Abolaji, T. O. (2020). Conceptual model for privacy-centric security engineering in digital and cloud computing systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(5), 371–389.
150) Okoruwa, P. O., Fadayomi, O., Abolaji, T. O., Edivri, J., Ogbole, J. I., & Akeju, B. (2024). Enterprise cybersecurity trends and threat evolution: Advances and emerging research directions. Global Multidisciplinary Perspectives Journal, 1(6), 194–204. https://doi.org/10.54660/GMPJ.2024.1.6.194-204
151) Okoruwa, P. O., Fadayomi, O., Bello, A. D., Elebe, O., Hammed, N. I., & Omoegun, G. O. (2025). An artificial intelligence-driven financial crime investigation framework for analyst decision support. International Journal of Multidisciplinary Research and Growth Evaluation, 6(6). https://www.allmultidisciplinaryjournal.com/article/5453/
152) Oteri, O., & Edivri, J. (2024). Applied performance optimization frameworks for managing high-demand enterprise technology programs and system rollouts. Gyanshauryam, International Scientific Refereed Research Journal, 7(1), 188–210.
153) Oteri, O., & Edivri, J. (2026). An operational reliability and service assurance framework for enterprise IT systems supporting large user populations. Gulf Journal of Engineering & Technology, 2(1), 1–19. https://doi.org/10.51594/gjet.v2i1.205
154) Ozowara, D. E., Adebayo, A., & Anunagba, C. O. (2022). A systematic review of cybersecurity investments and their impact on healthcare financial performance. Shodhshauryam, International Scientific Refereed Research Journal, 5(1), 404–427. https://doi.org/10.32628/SHISRRJ

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

155) Ozowara, D. E., Anunagba, C. O., & Adepoju, P. A. (2025). A review of ransomware economics and financial resilience strategies in hospital networks. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2284–2298. https://doi.org/10.62225/2583049X.2025.5.6.6052

156) Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of EMNLP, 1532–1543. https://doi.org/10.3115/v1/D14-1162

157) Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Proceedings of NAACL-HLT, 2227–2237.

158) Quainoo, R., Ogundapo, O., & Asiedu, W. A. (2024). Modeling the impact of impedance mismatch on wireless link performance: A framework for prototype development and system optimization. International Journal of Computer Science and Mathematical Theory, 10(2), 62–116. https://doi.org/10.56201/ijcsmt.v10.no2.2024.pg62.116

159) Quainoo, R., Ogundapo, O., & Asiedu, W. A. (2025a). Predicting throughput degradation in Wi-Fi 6 networks under varying channel conditions: A physical layer performance model. International Journal of Engineering and Modern Technology, 11(12), 205–264. https://doi.org/10.56201/ijemt.vol.11.no12.2025.pg205.264

160) Quainoo, R., Ogundapo, O., & Asiedu, W. A. (2025b). Test automation in wireless hardware engineering: A comprehensive review of scripting frameworks and instrument control strategies. International Journal of Engineering and Modern Technology, 11(10), 405–464. https://doi.org/10.56201/ijemt.vol.11.no10.2025.pg405.464 Predictive Analytics for Retail, Finance, and Operations

161) Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

162) Rahman, M. R., Mahdavi-Hezaveh, R., & Williams, L. (2020). A literature review on mining cyberthreat intelligence from unstructured texts. Proceedings of the IEEE International Conference on Data Mining Workshops, 516–525.

163) Ramnani, R. R., Shivaram, K., & Sengupta, S. (2017). Semi-automated information extraction from unstructured threat advisories. Proceedings of the 10th Innovations in Software Engineering Conference, 181–187.

164) Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP, 3982–3992. https://doi.org/10.18653/v1/D19-1410

165) Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

166) Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. Journal of Big Data, 7(1), 41.

167) Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423.

168) Smart, B. D. (2020). A quantitative cyber risk valuation model for board-level decision making in critical infrastructure. International Journal of Multidisciplinary Research and Growth Evaluation, 1(5), 984–991. https://doi.org/10.54660/.IJMRGE.2020.1.5.984-991

169) Smart, B. D. (2023). A predictive supply chain cyber risk intelligence framework for infrastructure and defense projects. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2905–2910. https://doi.org/10.62225/2583049X.2023.3.6.6228

170) Smart, B. D., & Jooda, D. (2023a). A unified multi-framework compliance optimization model integrating ISO 27001, NIST Cybersecurity Framework, and Cybersecurity Maturity Model Certification. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://doi.org/10.32628/IJSRCSEIT

171) Smart, B. D., & Jooda, D. (2023b). A continuous compliance and real-time audit readiness architecture for regulated enterprises. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2911–2916. https://doi.org/10.62225/2583049X.2023.3.6.6229

172) Smart, B. D., & Jooda, D. (2024a). A national-scale AI governance and risk control framework for critical infrastructure protection. Gyanshauryam, International Scientific Refereed Research Journal, 7(6), 361–371. https://doi.org/10.32628/GISRRJ

173) Smart, B. D., & Jooda, D. (2024b). A cybersecurity maturity acceleration model for critical infrastructure systems. Gyanshauryam, International Scientific Refereed Research Journal, 7(6), 348–360. https://doi.org/10.32628/GISRRJ

174) Smart, B. D., & Jooda, D. (2025). Resilient security architecture for operational technology environments under advanced persistent threat conditions: A layered defense model for critical infrastructure protection. International Journal of Engineering and Modern Technology, 11(11), 147–159. https://doi.org/10.56201/ijemt.vol.11.no11.2025.pg147.159

175) Smart, B. D., & Jooda, D. (2026). Integrated incident response governance model (IIRG): Unifying strategic leadership, cross-sector threat intelligence, and AI-augmented response for enterprise cyber resilience. Gulf Journal of Computer Sciences, 2(1), 1–15. https://doi.org/10.51594/gjcs.v2i1.210

176) Strom, B. E., Applebaum, A., Miller, D. P., Nickels, K. C., Pennington, A. G., & Thomas, C. B. (2018). MITRE ATT&CK: Design and philosophy. MITRE Corporation.

177) Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., & Zhang, J. (2023). Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. IEEE Communications Surveys & Tutorials, 25(3), 1748–1774.

178) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2021a). Advances in Demand Forecasting: Machine Learning Algorithms for Revenue Projection and Financial Planning Accuracy. Gyanshauryam, International Scientific Refereed Research Journal, 4(1), 282–317. https://doi.org/10.32628/GISRRJ120361

179) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2021b). Predictive Cash Flow Modeling in Retail: A Review of Advanced Forecasting and Working Capital Optimization. Shodhshauryam, International Scientific Refereed Research Journal, 4(3), 366–397. https://doi.org/10.32628/SHISRRJ214349

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

180) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2022a). Advances in Geospatial Analysis: Predictive Analytics Methods for Store Location Selection and Market Entry Return on Investment. International Journal of Marketing and Communication Studies, 6(2), 78–105. https://doi.org/10.56201/ijmcs.v6.no2.2022.pg78.105
181) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2022b). Algorithmic Process Optimization and Cost Reduction: A Review of Simulation Modeling and Financial Impact Assessment. Journal of Accounting and Financial Management, 8(8), 139–169. https://doi.org/10.56201/jafm.vol.8.no8.2022.p139.169
182) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2023). A Conceptual Framework for Retail Asset Valuation: Integrating Quantitative Pricing and Multi-Unit Financial Performance Prediction. Journal of Accounting and Financial Management, 9(12), 218–250. https://doi.org/10.56201/jafm.v9.no12.2023.pg218.250
183) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2024a). Advances in Supply Chain Resilience: Predictive Models for Vendor Risk Assessment and Procurement Cost Optimization. International Journal of Social Sciences and Management Research, 10(11), 525–551. https://doi.org/10.56201/ijssmr.v10.no11.2024.pg.525.551
184) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2024b). Real-Time KPI Tracking Systems: A Review of Automated Performance Monitoring and Data-Driven Decision Making. World Journal of Innovation and Modern Technology, 8(6), 247–281. https://doi.org/10.56201/wjimt.v8.no6.2024.pg247.281
185) Tonoyan, A., Dada, O., & Ayivi-Donkor, S. S. (2026). Conceptual Framework for Financial Brand Architecture Optimization: Portfolio Analysis and Economic Value Maximization. Gulf Journal of Advance Business Research, 4(3), 145–164. https://doi.org/10.51594/gjabr.v4i3.213
186) Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers & Security, 72, 212–233. https://doi.org/10.1016/j.cose.2017.09.001
187) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
188) Wagner, T. D., Mahbub, K., Palomar, E., & Abdallah, A. E. (2019). Cyber threat intelligence sharing: Survey and research directions. Computers & Security, 87, 101589. https://doi.org/10.1016/j.cose.2019.101589

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Corresponding Author: Chukwunenye Amadi