*Malaysian Journal of Computing*, 11 (2): 2667-2685, 2026
eISSN: 2600-8238

# TURNING ONTOLOGIES INTO VECTORS: EMBEDDING MALAY ONTOLOGICAL KNOWLEDGE FOR SMARTER MACHINE LEARNING

**Zayanah Zafirah Zulkipli[1], Ruhaila Maskat[2], Noor Hasimah Ibrahim Teo[3*], and Zhang Jie[4]**

*[1]Computer Science Department, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Shah Alam, Selangor, Malaysia*
*[2]Information Systems Department, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Shah Alam, Selangor, Malaysia*
*[3*]Computer Science Department, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Jasin, Melaka, Malaysia*
*[4]Faculty of Mathematics and Computer, Xinyu University, Xinyu, Jiangxi, China*
[1]zayanahzafirahaz@gmail.com, [2]ruhaila@fskm.uitm.edu.my, [3*]shimateo@uitm.edu.my, [4]zhangjie@xyc.edu.cn
*(* indicates the corresponding author)*

## ABSTRACT

*This conceptual study proposes a three-layer methodological framework for representing Malay ontological knowledge in continuous vector spaces and integrating ontology-derived representations with machine-learning models. The framework comprises an Ontology Layer, an Embedding Layer, and an Integration Layer. Its contribution is not a new ontology-embedding algorithm rather, it provides a Malay-oriented methodological organization that links ontology construction, embedding-method selection, symbolic-subsymbolic fusion, ontology-guided consistency checking, and downstream evaluation. In the present study, the ontology layer is represented through a prototype Malay linguistic ontology and its class hierarchy and knowledge graph, whereas the embedding, integration, consistency-checking, and downstream evaluation stages are proposed components for subsequent empirical validation. Existing approaches including RDF2Vec, OPA2Vec, OWL2Vec*, and EL Embeddings are used as methodological reference points because they differ in graph representation, lexical information, and preservation of logical semantics. The framework is intended to support future Malay NLP experiments in which ontology-enhanced models can be compared with equivalent models without ontology-derived features. Accordingly, claims concerning accuracy, explainability, semantic disambiguation, and model improvement are treated as hypotheses for future empirical assessment rather than demonstrated outcomes.*

**Keywords:** *Hybrid Intelligence, Malay Ontology Symbolic–Subsymbolic Integration, Natural Language Processing (NLP), Ontology Embedding.*

Received for review: 10-09-2026; Revised: 28-09-2026; Accepted: 29-09-2026; Published: 01-10-2026
DOI: 10.24191/mjoc.v11i2.12579

This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/3.0/).

## 1. Introduction

The convergence of machine learning and symbolic knowledge representation has garnered a lot of scholarly attention recently, especially when it comes to embedding structured ontologies into continuous vector spaces. In knowledge-intensive applications like the semantic web, bioinformatics, and information systems, formalized ontologies, explicit descriptions of domain ideas and relationships are essential (Chen *et al*., 2024). In order to facilitate scalable knowledge representation and reuse, it is critical to integrate disparate academic data with ontology alignment, as demonstrated by recent attempts toward unified research infrastructures (Wilinski, 2026). Extensive surveys also show that, in complicated domains, ontology embedding has emerged as a crucial method for connecting symbolic knowledge and machine learning models (Chen *et al*., 2025). To further preserve logical semantics during embedding construction, sophisticated embedding techniques as lattice-based ALC representations have also been put forth (Zhapa-Camacho and Hoehndorf, 2025).

However, traditional ontological reasoning approaches occasionally exhibit difficulties when used in large-scale machine learning applications, which often call for efficient vector-based representations and statistical learning paradigms (Chen *et al*., 2024; Xiong *et al*., 2023). In multilingual and culturally specific domains like Malay, where computational resources and embeddings are still under-studied, the methodological gap is particularly significant (Ahmad, 2022).

This gap is acutely felt in Malay Natural Language Processing (NLP), where the language's morphological complexity and informal social media registers (containing slang, *Bahasa SMS*, and "*Bahasa Rojak*") present major hurdles (Enjop *et al*., 2022). Historically, a common workaround has been translating low-resource Malay text into a resource-rich language like English using commercial Machine Translation (MT) like Google Translate (Enjop *et al*., 2022). However, this translation-first approach is highly problematic because MT regularly distorts semantic context and sentiment polarity (Enjop *et al*., 2022). For example, MT routinely mistranslates colloquial Malay pronouns such as "die" (meaning "he/she/it") into the English verb "die", introducing severe negative sentiment bias into downstream models (Enjop *et al*., 2022). Furthermore, commercial engines fail to resolve common colloquial short forms like "*smua*", "*hg*", "*tjk*", and "*mluat*", forcing translation tools to output uninformative, neutral scores (Enjop *et al*., 2022). This failure highlights the urgent need to model and embed Malay ontological structures directly, rather than relying on error-prone translation pipelines.

To address these linguistic barriers, this paper focuses on the conceptual problem of how a Malay-language ontology can be transformed into vector representations and subsequently connected to machine-learning models without discarding the structured semantic information encoded in the ontology. The aim is to conceptually propose and describe a methodological framework rather than to report a completed empirical evaluation. The framework therefore specifies the information flow from symbolic ontology construction to ontology embedding and hybrid integration, while identifying the mechanisms and evaluation procedures required for later implementation. The central challenge is to balance the semantic, hierarchical, and axiomatic structure of ontologies with the continuous numerical representations required by subsymbolic learning (Chen *et al*., 2018; Chen *et al*., 2024). Existing systems such as OWL2Vec* demonstrate that lexical, structural, and logical ontology content can be represented in continuous vector spaces (Chen *et al*., 2021); the present work uses such methods as reference approaches rather than claiming a new embedding algorithm.

In the context of Malay NLP, this builds directly upon structured lexical resources developed in prior literature. Early projects constructed foundational databases such as WordNet Bahasa which mathematically defines verbs, adverbs, and negation rules (Enjop *et al*., 2022) and RojakLex which utilizes valence shifters to parse mixed language registers (Enjop *et al*., 2022) to represent Malay semantic relationships. However, to create "smarter" machine learning systems, the next computational milestone is to transform these discrete,

symbolic taxonomies and ontologies into continuous vector embeddings, mapping hierarchical concepts into unified geometric spaces.

Accordingly, the contribution of this paper is conceptual and methodological. First, it organizes a Malay-oriented ontology-to-machine-learning pipeline into three explicit layers such as symbolic ontology representation, ontology embedding, and hybrid integration. Second, it clarifies how established ontology-embedding families can be selected according to the information that must be preserved, including graph topology, lexical annotations, hierarchy, and logical axioms. Third, it proposes an integration and ontology-consistency mechanism together with an evaluation plan for comparing ontology-enhanced models against equivalent baselines without ontology-derived representations. The novelty therefore lies in the combination and methodological organization of these components for Malay NLP, rather than in a newly invented embedding algorithm. The ontology prototype presented in this paper illustrates the ontology construction, embedding, integration, consistency, and downstream evaluation stages remain proposed and require empirical validation.

## 2. Literature Review

The conversion of symbolic ontologies into continuous vector representations has advanced significantly in recent years. Numerous embedding strategies have been put forth; these differ primarily in how they include lexical or metadata information while maintaining formal logical semantics. These methods are crucial for establishing a connection between machine learning models and symbolic knowledge representation, enabling the encoding of structured semantic information into vector spaces (Chen *et al*., 2021; Chen *et al*., 2024). The necessity for embedding techniques that can successfully balance semantic correctness, computational efficiency, and flexibility across many domains is further highlighted by recent studies (Chen *et al*., 2025).

Using logical axioms and annotations to create vector representations was the main focus of early ontology embedding research. Onto2Vec, one of the first methods, creates embeddings from ontology annotations and formal axioms to enable similarity-based prediction tasks, especially in biological applications (Smaili *et al*., 2018). This approach shows that logical reasoning-derived embeddings can facilitate supervised learning tasks like interaction prediction. Its use is, however, mainly limited to biological ontologies, which are very different from other domain-specific ontologies.

By combining formal axioms with informal textual data, such as labels, definitions, and synonyms, OPA2Vec expands on this concept. Additionally, it uses pre-trained Word2Vec models to integrate linguistic context into the embeddings (Smaili *et al*., 2019; Smaili *et al*., 2018). This method enhances performance in tasks like gene-disease prediction and protein interaction. However, there is a trade-off, whereas adding informal information improves semantic richness, it may lessen the maintenance of rigid logical linkages, particularly in applications that call for exact reasoning (Smaili *et al*., 2018). Given the scarcity of language resources in Malay ontologies, this problem is especially pertinent.

Hybrid methods that incorporate structural, lexical, and logical information have been created to overcome the shortcomings of previous approaches. One prominent example is OWL2Vec*, which combines lexical characteristics, logical axioms, and graph-based random walks into a single embedding framework (Chen *et al*., 2021). It is appropriate for multilingual applications due to its adaptable architecture, which enables various components to be modified individually. Nevertheless, these techniques still only approximate rather than completely preserve logical semantics.

Additionally, graph-based methods play a major role in ontology embedding. For example, RDF2Vec has demonstrated good performance in modeling knowledge graphs and uses random walk algorithms to capture graph structures (Portisch and Paulheim, 2024). In a

broader sense, knowledge graph reasoning has been thoroughly investigated, with research highlighting important issues like scalability, logical consistency, and inference efficiency (Tian *et al*., 2022; Zhang *et al*., 2022; Tachmazidis *et al*., 2018). Graph neural networks (GNNs) and graph learning techniques support this by offering strong tools for modeling relational data and deriving significant patterns from graph topologies (Zhou *et al*., 2018; Wu *et al*., 2020; Chami *et al*., 2020).

Embedding techniques that explicitly maintain logical semantics are the focus of an alternate approach. In order to approximate the compliance of logical constraints, EL Embeddings offer a geometric framework for modeling description logic EL++ ontologies in vector spaces (Kulmanov *et al*., 2019). These methods are especially helpful in situations when preserving logical consistency is crucial. However, they are mostly assessed in biomedical situations and are typically computationally demanding.

PO2Vec, which uses hierarchy-aware vector learning to more clearly encode ontological structures, represents another pertinent direction (Li *et al*., 2024). PO2Vec integrates hierarchical ontology structure, transitive reasoning, and embedding learning into a hybrid framework, in contrast to methods that mainly rely on random walks or lexical co-occurrence. As a result, structural semantic dependencies across ontology levels can be more effectively preserved in the final vector representations. Because Malay semantic systems frequently need machine-readable representations that preserve subclass relations, taxonomic order, and culturally grounded semantic differences, such a hierarchy-sensitive technique is especially pertinent for Malay ontology embedding.

The shortcomings of traditional embedding techniques have been brought to light by additional theoretical research. Many knowledge graph embeddings have been demonstrated to be incapable of correctly capturing intricate logical rules or relationships, which may result in a loss of semantic meaning (Gutiérrez-Basulto and Schockaert, 2018). Logic-based embedding approaches have been offered as a solution to this problem, allowing for efficient computation while preserving formal semantics (Abboud *et al*., 2022). These factors are particularly crucial for culturally grounded ontologies, such as those found in the Malay language, where maintaining semantic integrity is essential.

Strong contextual embedding methods that can support ontology-based representations have been made possible by recent developments in natural language processing. Large-scale pretraining has helped transformer models like RoBERTa and T5 perform well on a range of tasks (Liu *et al*., 2021; Raffel *et al*., 2021). Furthermore, Sentence-BERT uses Siamese network topologies to efficiently compute semantic similarity at the sentence level (Reimers and Gurevych, 2021).

In order to capture increasingly intricate linguistic patterns, neural embedding techniques have also developed. Word2Vec extensions make it possible to express sentences and phrases, which enhances their utility in subsequent applications (Mikolov *et al*., 2013b). These findings imply that semantic understanding can be improved by integrating ontology embeddings with contextual language models, especially in low-resource and multilingual environments.

Neuro-symbolic AI is the result of combining neural learning and symbolic reasoning. This paradigm improves interpretability and generalization capabilities by combining data-driven methods with logical inference (Xiong *et al*., 2023). The integration of logical reasoning into neural networks, enabling end-to-end learning of reasoning processes, is made possible by recent research on differentiable reasoning (Minervini *et al*., 2020; Marra *et al*., 2024).

In the meantime, relational data modelling still heavily relies on graph neural networks. Their expressive potential and limitations in representing intricate graph structures are the subject of ongoing research (Xu *et al*., 2019; Sato, 2020; Zhang *et al*., 2023). By enabling multi-hop reasoning and forecasting missing links within knowledge graphs, further methods like composition-based embeddings and knowledge graph completion improve these models.

Recent advancements between 2023 and 2026 show that knowledge graph reasoning and ontology embedding are still developing. In specialized domains like bioinformatics, domain-specific embedding techniques have proven to perform well (Li *et al*., 2024). The significance of creating embedding techniques that strike a compromise between semantic correctness, scalability, and adaptability is further highlighted by recent surveys (Chen *et al*., 2024).

Despite these developments, three gaps motivate the proposed framework. First, most established ontology-embedding methods have been developed and evaluated in English-language, biomedical, or general knowledge-graph settings, while their methodological application to Malay linguistic resources remains limited. Second, existing methods differ substantially in what they preserve such as RDF2Vec emphasises graph neighbourhoods, OPA2Vec combines formal and informal ontology content, OWL2Vec* combines structural, lexical, and logical views, and EL Embeddings explicitly model selected description-logic semantics. Consequently, no single method should be assumed to preserve all information required by a Malay NLP application. Third, evaluation must consider both downstream task performance and the extent to which ontology structure or semantics are retained. These gaps motivate a framework that separates ontology construction, embedding selection, integration, consistency checking, and evaluation as explicit methodological decisions.

Overall, the literature indicates that ontology-embedding approaches make different trade-offs among structural representation, lexical richness, logical preservation, and computational flexibility. The framework proposed in this study therefore does not replace RDF2Vec, OPA2Vec, OWL2Vec*, EL Embeddings, or related techniques. Instead, these methods can operate within the Embedding Layer according to task requirements. The distinctive contribution is a Malay-oriented end-to-end methodology that connects a structured linguistic ontology to embedding selection, hybrid fusion, ontology-guided validation, and a defined empirical evaluation strategy.

Table 1. Comparison of Ontology Embedding Methods

| Approach | Input Representation /Embedding Method | Axiom / Semantic Preservation | Integration / Target Task |
|---|---|---|---|
| RDF2Vec (Portisch & Paulheim, 2024) | RDF graph; random walks followed by Word2Vec-style learning | Primarily structural; does not guarantee logical-axiom preservation. | General knowledge-graph representation and downstream ML. |
| OPA2Vec (Smaili *et al*., 2019) | Formal ontology axioms plus labels, definitions and other annotations; Word2Vec-based learning | Includes formal and informal ontology content, but does not guarantee complete logical preservation. | Similarity-based prediction, originally demonstrated in biomedical domains. |
| OWL2Vec* (Chen *et al*., 2021) | OWL ontology represented through structural, lexical and logical document views | Captures logical information approximately together with structure and lexical content. | Ontology embeddings for downstream machine-learning tasks. |
| EL Embeddings (Kulmanov *et al*., 2019) | EL++ ontology mapped to geometric vector-space constraints | Explicitly designed to preserve selected EL++ logical semantics. | Logic-sensitive ontology representation and prediction. |
| Proposed Malay framework | Malay ontology prototype → selected embedding method → contextual representation → hybrid fusion | Preservation depends on selected embedding method; ontology rules are additionally used for proposed consistency checking. | Malay NLP; framework-level integration, baseline comparison, ablation and semantic evaluation. |

Table I positions the proposed framework relative to representative ontology-embedding approaches. RDF2Vec primarily captures graph neighbourhoods through random walks; OPA2Vec combines formal ontology content with textual annotations; OWL2Vec* combines structural, lexical, and logical views; and EL Embeddings emphasise geometric preservation of EL++ semantics. In contrast, the proposed Malay framework is not an alternative embedding algorithm. It is an end-to-end methodological architecture in which an appropriate embedding method is selected and then connected to a Malay NLP model through an explicit fusion and ontology-consistency stage.

This distinction is important because the proposed contribution operates at the framework level. The embedding technique is treated as a configurable component, while the framework specifies the inputs to be represented, the criteria for selecting an embedding family, the mechanism for combining ontology-derived and contextual representations, and the evaluation required to determine whether ontology information provides measurable benefit.

Thus, the proposed framework is method-agnostic at the embedding stage but method-specific at the architectural level which requires a symbolic ontology representation, an explicitly selected ontology-embedding procedure, a defined hybrid fusion mechanism, ontology-guided consistency checking, and comparative evaluation against a non-ontology baseline.

## 3. Theoretical and Conceptual Framework

The theoretical framework of this conceptual article integrates three primary perspectives which is symbolic knowledge representation theory, hybrid symbolic–subsymbolic learning, and distributional semantics. These theories work together to support the conversion of Malay ontological knowledge into vector representations that machine learning (ML) systems can use efficiently.

### 3.1 Symbolic Knowledge Representation Theory

Ontologies are formalized conceptual schemas that represent entities, relationships, and axioms using logic-based formalisms like Web Ontology Language (OWL) are based on symbolic knowledge representation, as defined in classical artificial intelligence. Machines can reason deductively over organized knowledge through ontologies (Gruber, 1993). However, symbolic systems by themselves struggle with scalability and ambiguity when exposed to real-world data or plain language. Though their strict logical syntax restricts integration with deep learning models that depend on continuous numerical input, symbolic ontologies in the Malay context encode culturally distinctive language and conceptual structures.

### 3.2 Distributional Semantics and Vector Space Theory

Vector embeddings are theoretically based on the distributional semantics principle, which states that "you shall know a word by the company it keeps" (Firth, 1957). Words that are semantically similar are positioned in close proximity in high-dimensional space in representations learned by models like Word2Vec which is designed for the "efficient estimation of word representations in vector space" (Mikolov *et al*., 2013), where each word is represented as a dense, low-dimensional vector in a continuous space. Word2Vec have also been extended to capture compositional semantics of phrases and sentences, improving representation quality in downstream tasks (Mikolov *et al*., 2013b). Next model is Global Vectors for Word Representation (GloVe) is a count-based model that uses co-occurrence matrix which is designed to constructs word embeddings by capturing "global corpus statistics" and encoding them into a vector space representation (Pennington *et al*., 2014). Another model is fastText which is Word2Vec extended by incorporating subword (character n-gram) information. Words are represented as combination of smaller units and embeddings are learned by modelling character n-grams in fastText model (Bojanowski *et al*., 2016).

Despite their success in capturing latent language links, Word2Vec, GloVe and fastText embeddings are still only statistical in nature and do not explicitly convey conceptual meaning (Mikolov *et al*., 2013). By recording ontological semantics as geometric patterns in continuous space, embedding ontologies into vector spaces expands this idea beyond words to concepts and relations (Bouraoui *et al*., 2022).

### 3.3 Hybrid Symbolic-Subsymbolic Learning

The strengths of neural learning and symbolic reasoning are combined in the new idea of hybrid intelligence, also known as neuro-symbolic integration (Besold *et al*., 2017). It suggests that ML models can reason more intelligently by utilizing both data-driven generalization and rule-based inference when symbolic knowledge, such as ontological axioms, is embedded into neural vector spaces. Ontology embedding systems such as OWL2Vec* (Chen *et al*., 2021), which maintain ontological structure while transforming lexical elements and logical axioms into embeddings suitable for downstream machine learning applications, best illustrate this hybrid paradigm.

### 3.4 Application to the Research Problem

The theoretical perspectives above are operationalised as a three-layer conceptual and methodological framework for incorporating Malay ontological information into machine-learning workflows. Figure 1 presents the proposed end-to-end pipeline. The Ontology Layer is represented in this study by a prototype Malay linguistic ontology, including the class hierarchy and knowledge graph shown in Figures 2 and 3. The Embedding Layer and Integration Layer define the proposed procedures for converting this symbolic resource into vectors and combining those vectors with task-specific contextual representations. These latter layers, including ontology-guided consistency checking and downstream evaluation, are not presented as completed experiments in this paper. This explicit distinction between the implemented prototype and proposed components prevents the conceptual architecture from being interpreted as an empirically validated system.

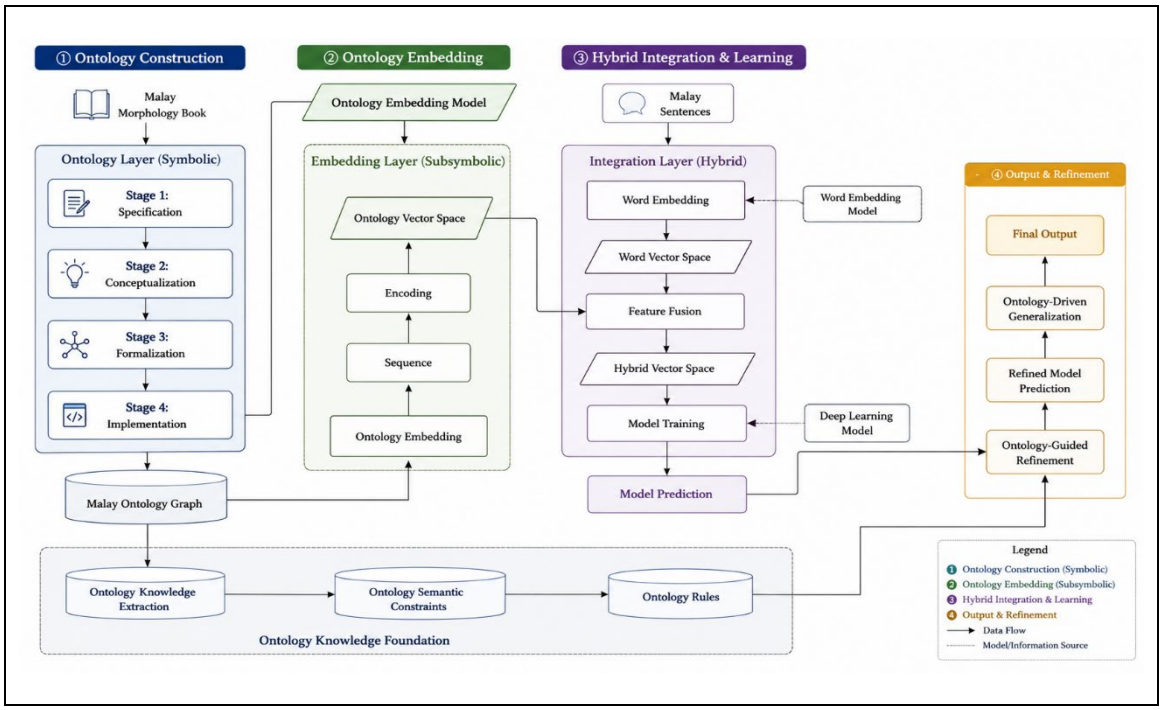



Figure 1. Full pipeline of the framework

#### 3.4.1 Ontology Layer (Symbolic)

**Input:** The Ontology Layer provides the symbolic foundation of the framework. In the present study, a prototype Malay linguistic ontology was constructed from Malay grammatical and morphological concepts using *Tatabahasa Dewan*, Third Edition (Nik Safiah Karim, *et al*., 2008), as the principal linguistic reference. The source provides established descriptions of

Malay grammatical structures and categories. Relevant linguistic concepts were systematically identified from the reference and represented as ontology classes, subclasses, properties, relationships, and axioms. These structured elements constitute the symbolic input for the subsequent embedding layer.

**Process:** Ontology construction follows four stages such as specification, conceptualization, formalization, and implementation. During specification, the purpose, scope, users, and intended NLP use are defined. During conceptualization, Malay linguistic concepts are organized into taxonomies and semantic relationships. During formalization, these concepts and constraints are expressed using OWL-compatible classes, properties, and axioms. During implementation, the formal representation is instantiated as a machine-readable knowledge graph. This process constitutes the implemented prototype component of the present conceptual study.

**Stage 1 - Specification**: The ontology is scoped to Malay linguistic knowledge relevant to semantic representation and future NLP applications, including POS-related categories. Intended users include NLP researchers and developers of Malay language-processing systems. The output is a specification describing the ontology purpose, domain boundaries, intended concepts, and planned downstream use.

**Stage 2 - Conceptualization**: Core linguistic classes include *kata_nama* (noun), *kata_adjektif* (adjective), and *kata_kerja* (verb), with subclass structures such as *kata_ganti_nama* ⊆ *kata_nama*. Lexical-semantic relations such as synonymy, antonymy, and part-whole relations can also be represented where supported by the linguistic source. Properties such as hasEmotion, hasIntensity, and relatesToCulture are retained only where their meaning and domain/range are explicitly defined. Cultural semantics are treated as an extension requiring evidence from appropriate Malay cultural or lexical sources rather than being inferred solely from grammatical categories.

**Stage 3 - Formalization**: The conceptual model is expressed using the Web Ontology Language (OWL). Classes, subclass relations, object/data properties, domain and range restrictions, and applicable axioms are formally specified. This stage provides the machine-readable semantics required for later reasoning and embedding.

**Stage 4 - Implementation**: The formalised ontology is instantiated as a machine-readable OWL/RDF knowledge graph. Figures 2 and 3 are outputs of this ontology prototype and therefore represent the implemented component of the framework rather than results from the embedding or machine-learning stages.

**Ontology validation**: Because ontology quality directly affects subsequent embeddings, validation is treated as a required methodological step. Structural validation should check class hierarchy, duplicate concepts, domain/range declarations, and relation connectivity; logical validation should use an OWL reasoner to identify inconsistent or unsatisfiable classes; and semantic validation should involve review by Malay language/domain experts against the linguistic source. Where these checks have not yet been completed for the prototype, they are requirements for the empirical implementation rather than claimed completed evaluations.

- Class: *kata_adjektif_bentuk* ⊆ *kata_adjektif*
- Property: expressesEmotion → *gembira*

**Output:** The output is a structured Malay linguistic knowledge graph that can serve as the symbolic input to the proposed Embedding Layer. In the current paper, this ontology prototype demonstrates the feasibility of representing Malay grammatical concepts in machine-readable form. Broader cultural concepts, culturally dependent meanings, and usage-specific relations should be added only when supported by appropriate Malay cultural or lexical evidence.

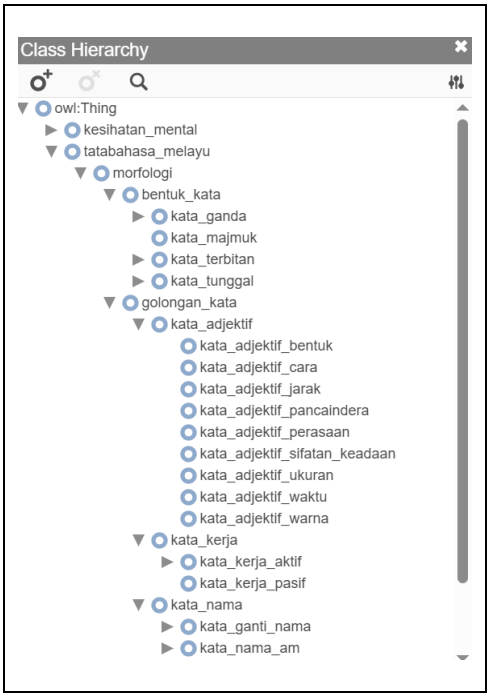



Figure 2. Ontology class hierarchy

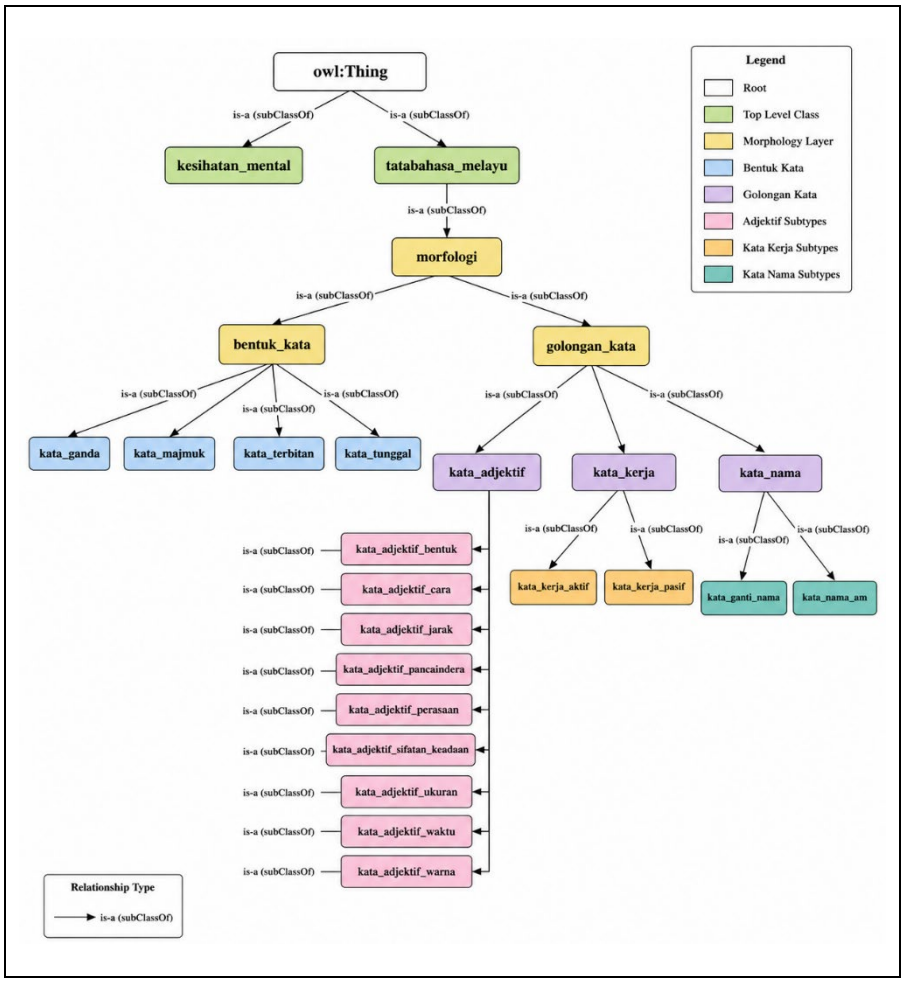



Figure 3. Ontology knowledge graph

### 3.4.2 Embedding Layer (Subsymbolic)

**Input:** The proposed Embedding Layer receives the OWL/RDF ontology graph, including triples, class labels, hierarchy information, axioms, and available lexical metadata. In this paper, this layer is specified conceptually; no single embedding algorithm is claimed as the implemented method.

**Process:** The framework is method-agnostic with respect to ontology embedding. The embedding method should be selected according to the information that must be preserved and the downstream task. RDF2Vec is suitable when graph neighbourhood and structural proximity are primary; OPA2Vec is relevant when formal axioms and textual annotations should both contribute; OWL2Vec* is appropriate when structural, lexical, and logical views are jointly required; and EL Embeddings are relevant when preservation of EL++ logical semantics is a central requirement. Selection should therefore consider ontology expressivity, availability of lexical annotations, need for axiom preservation, computational cost, vector dimensionality, and compatibility with the target ML architecture.

- For the first empirical validation, OWL2Vec* is proposed as a primary candidate because its multi-view treatment can represent structural, lexical, and logical ontology information. RDF2Vec can serve as a structure-focused comparator, while an appropriate logic-oriented method may be used where formal axiom preservation is prioritised. This recommendation is a future experimental design choice and is not presented as an implemented result in the current conceptual paper.
- Embedding parameters to be reported in future implementation include the selected algorithm, ontology input version, vector dimensionality, walk strategy or logical constraints where applicable, training objective, number of epochs, context/window parameters, random seed, and software implementation. Reporting these settings is necessary for reproducibility.
- Terminology: In this paper, ontology embedding refers to the process or method used to transform symbolic ontology content into a continuous vector space. Ontology embeddings, or ontology-derived vectors, refer to the resulting numerical representations of ontology entities or concepts. These terms are used consistently throughout the revised manuscript.

Expected representation: A successful embedding should place semantically related concepts in meaningful geometric relationships while retaining the ontology information targeted by the selected method. This is an expected property to be tested rather than an outcome demonstrated in the present study.

**Output:** The proposed output is a vector representation for ontology entities that can be passed to the Integration Layer. The numerical vectors shown in Figure 5 are illustrative representations of the expected output format and are not used as empirical evidence of framework effectiveness in this conceptual study.

- *kata_adjektif* → [0.24, 0.58, 0.11, ...]
- *kata_sifat* → [0.25, 0.59, 0.09, ...]

Figure 5 therefore illustrates how an ontology class may be represented numerically after an embedding procedure. It should not be interpreted as a reported experimental result. A future implementation must report the exact embedding algorithm, dimensionality, parameters, ontology version, and training procedure used to generate the vectors.

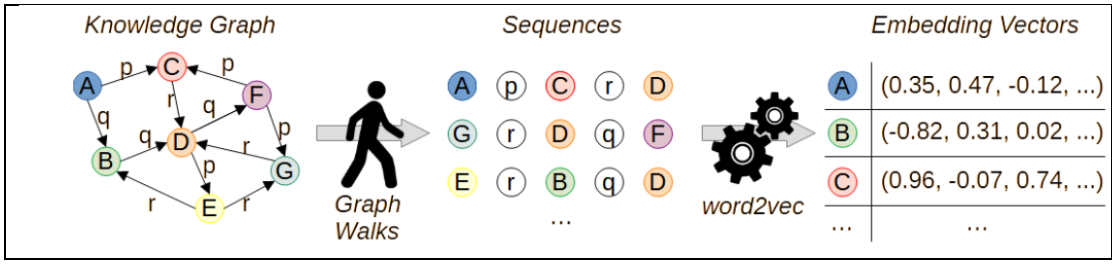



Figure 4. Conceptual illustration of a graph-walk-based ontology embedding process (RDF2Vec)

### 3.4.3 Integration Layer (Hybrid)

**Input:** The proposed Integration Layer receives (i) ontology-derived vectors from the Embedding Layer and (ii) contextual representations generated from a task-specific Malay dataset, such as a POS-tagged corpus. This layer has not been empirically evaluated in the present paper.

**Process:** The proposed fusion mechanism aligns each task token or concept with a corresponding ontology concept where a valid mapping exists. The contextual representation generated by a word-embedding, BiLSTM, or Transformer encoder is combined with corresponding ontology-derived vector. Before integration, both representations are projected into compatible dimensions to ensure that they can be processed jointly. Two alternative fusion strategies are proposed. The first uses concatenation, in which the projected contextual and ontology-derived representations are joined into a single feature representation before being passed to the task-specific prediction layer. This strategy provides a simple and transparent baseline for ontology integration. The second uses gated fusion, in which a learnable gating mechanism dynamically determines the relative contribution of contextual and ontology-derived information for each instance. A higher contextual weight causes the model to rely more strongly on information learned from the linguistic context, whereas a higher ontology weight increases the contribution of structured semantic knowledge. The weighted representations are then combined into a single fused representation and passed to the prediction layer. These two mechanisms are proposed for future implementation and empirical comparison and have not been evaluated in the present conceptual study.

- Step 1 - Mapping and alignment: task tokens or labels are mapped to ontology entities using an explicit lookup or mapping table. Unmapped items retain only their contextual representation.
- Step 2 - Feature fusion: contextual and ontology-derived representations are combined using concatenation as the initial baseline and, subsequently, a learnable gated-fusion mechanism.
- Step 3 - Ontology-guided consistency checking: after the model produces a probability distribution over candidate labels, predictions are checked against applicable ontology constraints such as class membership, subclass relations, domain/range restrictions, and explicitly encoded incompatibilities. A prediction that violates a relevant constraint can be flagged or reconsidered using the highest-probability ontology-compatible alternative. This is proposed as a post-prediction validation mechanism; it is not claimed to have been implemented or evaluated in the current study.

The integration stage is therefore designed to test whether symbolic knowledge provides information beyond contextual learning alone. Its effectiveness must be established empirically through baseline comparison and ablation rather than assumed from the architecture.

**Output:** The proposed Integration Layer is expected to produce task predictions informed by both contextual representations and ontology-derived features. Potential benefits include improved handling of semantically ambiguous or rare expressions and more ontology-

compatible predictions; however, these are hypotheses to be evaluated and are not reported as demonstrated improvements in this conceptual paper.

- Improved semantic disambiguation, particularly for polysemous or context-dependent Malay expressions.
- Enhanced processing of rare or unseen terms through ontology-driven generalization.
- Ontology-consistent predictions that ensure logical coherence in output labels or classifications.

Taken together, the proposed integration procedure provides a testable design for combining Malay linguistic knowledge with statistical learning. Its value will be determined by whether the ontology-enhanced configuration outperforms or otherwise complements an equivalent non-ontology baseline while preserving relevant ontology semantics.

```
Class: predikat_frasa_kerja_asal_berimbuhan_tidak_transitif
Embedding (first 10 elements): [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.] ...
Embedding shape: (300,)

Class: kata_adjektif_bentuk
Embedding (first 10 elements): [-0.00861641  0.05661298 -0.00078554  0.01287758  0.00695384 -0.0332093
  0.06033548  0.11609477  0.03818746 -0.02873828] ...
Embedding shape: (300,)

Class: kata_ganti_nama_diri_tanya
Embedding (first 10 elements): [ 0.01584111  0.17388847 -0.05631179  0.07615396 -0.09415692 -0.11982317
  0.13325244  0.41844654  0.09298959 -0.12226955] ...
Embedding shape: (300,)
```

Figure 5. Illustrative format of ontology-derived vectors (conceptual example; not an empirical evaluation result)

The three-layer design therefore specifies how symbolic Malay linguistic knowledge could be connected to subsymbolic learning without implying that downstream performance has already been demonstrated. The framework's contribution is the explicit methodological separation of ontology construction, embedding selection, hybrid fusion, consistency checking, and evaluation.

## 4. Proposed Evaluation and Validation Framework

As the present study proposes a conceptual framework, empirical evaluation of the complete framework has not yet been conducted. Therefore, this section defines a propFosed evaluation and validation strategy for future implementation. Part-of-Speech (POS) tagging is proposed as the initial Malay NLP task because the developed ontology prototype contains grammatical concepts and categories that can support the representation of Malay linguistic knowledge.

The proposed evaluation will compare several model configurations to determine the contribution of ontology-derived knowledge. First, a baseline model without ontology integration will be established using a suitable contextual model such as BiLSTM or a Transformer-based model. Second, an ontology-enhanced model using feature concatenation will combine contextual representations with ontology-derived representations. Third, a gated-fusion model will dynamically control the contribution of contextual and ontological information. Where ontology consistency checking is implemented, an additional comparison will examine the ontology-enhanced model with and without ontology-based post-prediction validation. This comparison provides an ablation strategy for determining whether changes in model performance can be associated with ontology integration and its individual components.

The downstream models will be evaluated using accuracy, precision, recall, and F1-score on the same held-out test set. In addition to task-level performance, the ontology and its derived representations should be evaluated independently. Ontology validation will examine

logical consistency and identify unsatisfiable classes using an OWL reasoner, while expert or linguistic validation will assess whether selected classes, relationships, and axioms appropriately represent the Malay linguistic knowledge derived from the reference source. Embedding evaluation will examine whether semantically related or hierarchically connected ontology concepts remain appropriately represented in the resulting vector space.

To ensure a fair comparison, all experimental configurations should use the same ontology version, dataset partition, preprocessing procedure, and evaluation measures. Repeated experimental runs and statistical significance testing may also be employed to determine whether observed differences are consistent rather than caused by random model variation. The primary objective of this evaluation is therefore to determine whether the incorporation of ontology-derived knowledge produces measurable differences compared with an equivalent model that relies only on contextual information. As these procedures constitute the proposed empirical validation protocol, their results will be reported in future work rather than in the present conceptual study.

## 5. Discussion

The principal contribution of the proposed framework is conceptual since it makes explicit the methodological path from a Malay linguistic ontology to a machine-learning representation. Rather than introducing a new embedding algorithm, the framework separates decisions that are often conflated in ontology-enhanced learning: what symbolic knowledge is represented, which ontology information an embedding method should preserve, how ontology-derived and contextual representations are fused, how ontology constraints can be checked, and how the resulting system should be compared with a non-ontology baseline. This organisation provides a testable basis for future Malay NLP research.

Relative to RDF2Vec, OPA2Vec, OWL2Vec*, and EL Embeddings, the framework operates at a different level of abstraction. Those approaches provide specific mechanisms for constructing vector representations from graph structure, ontology axioms, lexical annotations, or logical constraints. The proposed framework treats such techniques as interchangeable candidates within its Embedding Layer and extends the methodological scope to include Malay ontology construction, task-level fusion, ontology-guided consistency checking, and comparative evaluation. Its Malay contribution is therefore the application-oriented organisation of these components around a low-resource linguistic ontology, not a claim that the underlying embedding principles are new.

The framework also offers a potential basis for traceability because ontology-derived features can, in principle, be linked back to named ontology concepts and relations. However, the present study does not implement or evaluate a dedicated explainable-AI method. Consequently, explainability is not claimed as a demonstrated property. Future work should evaluate whether model decisions can be meaningfully traced to ontology concepts, for example through feature attribution, concept-level analysis, or rule-based explanations, and should assess the usefulness of those explanations to human evaluators.

The Malay ontology prototype currently provides stronger evidence for linguistic grounding than for cultural grounding. Its implemented examples are primarily grammatical categories such as *kata_nama*, *kata_kerja*, and *kata_adjektif*. Culturally dependent concepts, code-switching practices such as *Bahasa Rojak*, culturally specific lexical meanings, honorific usage, or other culturally conditioned relations could be incorporated in future extensions, but they require appropriate linguistic or cultural sources and expert validation. The revised framework therefore treats cultural representation as an intended extension rather than an empirically established property of the current prototype.

However, several limitations remain. First, ontology embeddings may approximate rather than fully preserve complex logical semantics, and the degree of preservation depends

on the selected method. Second, Malay resources, including mature ontologies and annotated corpora, remain comparatively limited, which may constrain large-scale implementation. Third, the ontology prototype requires documented structural, logical, and expert validation before it can serve as a validated knowledge resource. Fourth, the proposed fusion and consistency mechanisms have not yet been empirically tested. Finally, no claims of improved accuracy, interpretability, explainability, or semantic disambiguation can be confirmed until the framework is compared experimentally with appropriate baselines.

Despite these limitations, the framework provides a concrete research design for investigating ontology-enhanced Malay NLP. Candidate applications include POS tagging, semantic classification, named-entity recognition, and sentiment analysis, but each application should define its ontology mapping, baseline model, embedding configuration, fusion strategy, and evaluation criteria separately. The conceptual contribution is therefore a structured basis for experimentation rather than evidence that ontology integration necessarily improves a downstream model.

Overall, the framework reframes Malay ontology embedding as an end-to-end methodological problem involving knowledge modelling, representation learning, integration, constraint checking, and evaluation. This perspective provides a foundation for subsequent empirical studies while keeping the present paper's claims aligned with its conceptual scope.

## 6. Future Work

The first empirical validation should operationalise the framework on a specific Malay NLP task. POS tagging is proposed as the initial task because it aligns directly with the grammatical classes represented in the ontology prototype. A baseline BiLSTM or Transformer-based tagger without ontology-derived features should first be established. The same architecture should then be evaluated with ontology-derived vectors, initially using concatenation and subsequently gated fusion. OWL2Vec* is a suitable primary embedding candidate because it can combine structural, lexical, and logical ontology views, while RDF2Vec can provide a structure-oriented comparator. All configurations should use the same training, validation, and test partitions.

Evaluation should combine downstream task metrics (accuracy, precision, recall, and F1-score) with ontology-oriented measures. These should include logical consistency checks using an OWL reasoner, structural or hierarchy-preservation analysis, and expert assessment of selected Malay semantic relations. Ablation experiments should compare the baseline, ontology-enhanced model, and ontology-enhanced model with consistency checking. Repeated runs and appropriate statistical tests should be considered to determine whether observed differences are reproducible.

Subsequent work can extend the ontology beyond its current grammatical emphasis by incorporating culturally dependent Malay concepts from documented linguistic and cultural sources and validating them with domain experts. Explainability can then be investigated explicitly by tracing predictions to ontology concepts or rules and evaluating whether those explanations are understandable and useful. After validation on Malay, the same methodological framework may be examined in other low-resource Southeast Asian languages, with careful adaptation rather than assuming direct transferability.

## 7. Conclusion

This paper presents a conceptual and methodological framework for connecting Malay ontological knowledge with machine-learning representations through three layers: Ontology, Embedding, and Integration. The contribution is not a new ontology-embedding algorithm. Instead, the framework organises ontology construction, embedding-method selection, symbolic-subsymbolic fusion, ontology-guided consistency checking, and comparative evaluation into an end-to-end methodology for Malay NLP. The current study includes a prototype ontology layer, while the embedding, integration, consistency, and downstream evaluation components remain proposed for empirical validation. Existing approaches such as RDF2Vec, OPA2Vec, OWL2Vec*, and EL Embeddings can be selected within the Embedding Layer according to the structural, lexical, or logical information that a task requires. Potential benefits related to semantic disambiguation, accuracy, traceability, and explainability therefore remain hypotheses rather than demonstrated findings. Future work will test these hypotheses using a Malay NLP task, an equivalent baseline without ontology-derived features, controlled ablation experiments, ontology validation, and both task-level and semantic evaluation measures. In this way, the framework provides a clearly scoped foundation for investigating ontology-enhanced learning in a low-resource Malay setting without overstating empirical conclusions.

**Acknowledgement**

We would like to sincerely thank Universiti Teknologi MARA (UiTM) for its institutional assistance and ongoing academic supervision during the creation of this conceptual paper. Additionally, we are appreciative of the Faculty of Computer and Mathematical Sciences for its scholarly materials, helpful resources, and encouraging research atmosphere, all of which enabled us to investigate ontology and machine learning integration for the Malay language. We sincerely thank our research peers and colleagues for their insightful comments and thought-provoking conversations, which strengthened the study's theoretical and conceptual underpinnings. Lastly, we would like to express our gratitude to our families and mentors for their constant support and inspiration, which motivated us to finish this work on incorporating Malay ontological information for more sophisticated machine learning systems.

**Funding**

The authors received no specific funding for this work.

**Author Contribution**

Author 1, Author 2, Author 3 and Author 4 were actively involved in the conceptualisation, literature review, framework development, discussion, and writing of this paper. All authors reviewed and approved the final version of the manuscript.

**Declaration of Generative AI in the Writing Process**

During the preparation of this work, the author(s) ChatGPT (OpenAI) for language refinement and writing assistance, and QuilBot for paraphrasing and improving sentence clarity. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

**Data Availability/Supplementary Materials**

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

**Ethics Statement**

The authors declare that this research did not involve human or animal subjects. All experimental procedures were performed following the institutional Safety, Health, and Environmental (HSE) protocols of Universiti Teknologi MARA.

**Conflict of Interest**

The authors have no conflicts of interest to declare.

**References**

Abboud, R., *et al*. (2022). Logic embedding for knowledge graphs. *Artificial Intelligence*, 307.

Abd Rahman, H., & Pang, Y. (2026). A Conceptual Framework for Robust Few-Shot Learning: Integrating Unbalanced Optimal Transport and Self-Supervised Transformer Representations. *Malaysian Journal of Computing, 11*(1), 2378-2390.

Ahmad, M. (2022). Developing culturally sensitive AI models for multilingual contexts: A Southeast Asian perspective. *Journal of Artificial Intelligence Research, 73*(4), 220–237.

Besold, T. R., d'Avila Garcez, A., Bader, S., Bowman, H., Domingos, P., Hitzler, P., & Kuehnberger, K.-U. (2017). Neural-symbolic learning and reasoning: A survey and interpretation.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information.

Bouraoui, Z., Gutiérrez-Basulto, V., & Schockaert, S. (2022). Integrating ontologies and vector space embeddings using conceptual spaces. OASIcs – OpenAccess Series in Informatics, 3, 1–14.

Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C., & Murphy, K. (2022). Machine learning on graphs: A model and comprehensive taxonomy. *Journal of Machine Learning Research, 23*(89), 1–64.

Chen, J., Hu, P., Jimenez-Ruiz, E., Holter, O. M., Antonyrajah, D., & Horrocks, I. (2018). Ontology-based knowledge graph embedding. In *Proceedings of the International Semantic Web Conference (ISWC)*.

Chen, J., Hu, P., Jimenez-Ruiz, E., Holter, O. M., Antonyrajah, D., & Horrocks, I. (2021). OWL2Vec: Embedding of OWL ontologies. *Machine Learning, 110*(7), 2043–2071.

Chen, J., Mashkova, O., Zhapa-Camacho, F., Hoehndorf, R., He, Y., & Horrocks, I. (2025). Ontology embedding: A survey of methods, applications and resources. *IEEE Transactions on Knowledge and Data Engineering, 37*(7), 4193–4212.

Enjop, V., Adnan, R., Jamil, N., Ahmad, S., Zainol, Z., & Ahmad, S. A. (2022). Does Google Translate Affect Lexicon-Based Sentiment Analysis of Malay Social Media Text?. *Malaysian Journal of Computing, 7*(2), 1236-1249.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In Studies in linguistic analysis (pp. 1–32). Oxford: Blackwell.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 199–220.

Gutiérrez-Basulto, V., & Schockaert, S. (2018). From knowledge graph embedding to ontology embedding? An analysis of the compatibility between vector space representations and rules. In *Proceedings of the Sixteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2018)* (pp. 379–388). AAAI Press.

Ibrahim, N., Ishak, U. M., Ali, N. N. A., & Shaadan, N. (2024). Machine Learning-Based Approaches for Credit Card Debt Prediction. *Malaysian Journal of Computing, 9*(1), 1722-1751.

Karim, N. S., Onn, F. M., Musa, H. H., & Mahmood, A. H. (2008). Tatabahasa Dewan (3rd ed.). Dewan Bahasa dan Pustaka.

Kulmanov, M., Wang, L.-W., & Hoehndorf, R. (2019). EL Embeddings: Geometric construction for EL++ ontologies. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 6103–6109).

Li, W., Wang, B., Dai, J., Kou, Y., Chen, X., Pan, Y., Hu, S., & Xu, Z. Z. (2024). Partial order relation–based gene ontology embedding improves protein function prediction. *Briefings in Bioinformatics, 25(*2), bbae077.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2021). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.

Marra, G., Dumančić, S., Manhaeve, R., & De Raedt, L. (2024). From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence, 328*, 104062.

Manap, K. A., Ibrahim, M. S., & Ibrahim, N. Z. M. (2020). Instrument validation to determine the relationship between social media acceptance and Cuti-Cuti 1 Malaysia Facebook user engagement. *Malaysian Journal of Computing, 5*(1), 480–494

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS 2013)*, 26, 3111–3119.

Minervini, P., Riedel, S., Stenetorp, P., Grefenstette, E., & Rocktäschel, T. (2020). Learning reasoning strategies in end-to-end differentiable proving. *Proceedings of the 37th International Conference on Machine Learning (ICML 2020),* 119, 6938–6949.

Portisch, J., & Paulheim, H. (2024). The RDF2vec family of knowledge graph embedding methods: An experimental evaluation of RDF2vec variants and their capabilities. *Semantic Web, 15*(3), 845–876.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2021). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Sato, R. (2020). A survey on the expressive power of graph neural networks. arXiv preprint arXiv:2003.04078

Shafie, S. B., Ooi, S. P., & Khaw, K. W. (2023). Prediction of Employee Promotion Using Hybrid Sampling Method with Machine Learning Architecture. *Malaysian Journal of Computing, 8*(1), 1264-1286.

Smaili, F. Z., Gao, X., & Hoehndorf, R. (2018). Onto2Vec: Joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics, 34*(13), i52–i60.

Smaili, F. Z., Gao, X., & Hoehndorf, R. (2019). OPA2Vec: Combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics, 35*(12), 2133–2140.

Tian, L., Zhou, X., Li, Y., & colleagues. (2022). Knowledge graph and knowledge reasoning: A systematic review. *Journal of Electronic Science and Technology, 20*(2), 100159. https://doi.org/10.1016/j.jnlest.2022.100159

Tachmazidis, I., Antoniou, G., Batsakis, S., Urbani, J., & Zhou, Z. (2018). A survey of large-scale reasoning on the Web of data. *The Knowledge Engineering Review, 33*, e21. https://doi.org/10.1017/S0269888918000255

Wilinski, J. (2026). The science data lake: A unified open infrastructure integrating scholarly sources with ontology alignment. arXiv.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(1), 4–24.

Xiong, F., Zhang, H., & Chen, L. (2023). Hybrid symbolic and neural representation learning for knowledge integration. *Journal of Artificial Intelligence Research, 76*(2), 455–478.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? *International Conference on Learning Representations (ICLR 2019).*

Zhang, W., Chen, J., Li, J., Xu, Z., Pan, J. Z., & Chen, H. (2022). Knowledge graph reasoning with logics and embeddings: Survey and perspective. arXiv preprint arXiv:2202.07412.

Zhang, B., Fan, C., Liu, S., Huang, K., Zhao, X., Huang, J., & Liu, Z. (2023). The expressive power of graph neural networks: A survey. arXiv preprint arXiv:2308.08235.

Zhapa-Camacho, F., & Hoehndorf, R. (2025). Lattice-based ALC ontology embeddings with saturation.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI Open, 1, 57–81.