IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 November 2023, accepted 1 December 2023, date of publication 5 December 2023,
date of current version 13 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3339552

RESEARCH ARTICLE

Semi-Supervised Bootstrapped Syntax-Semantics-Based Approach for Agriculture Relation Extraction for Knowledge Graph Creation and Reasoning

G. VEENA[1], DEEPA GUPTA[2], AND VANI KANJIRANGAT[3]
[1]Department of Computer Science and Applications, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, Kollam 690525, India
[2]Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Bengaluru 560035, India
[3]Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA USI/SUPSI), 6962 Viganello, Switzerland

Corresponding author: Deepa Gupta (g_deepa@blr.amrita.edu)

ABSTRACT We propose a novel approach that uses semi-supervised learning to extract triplets from domain-specific texts and create a Knowledge Graph (KG), with a focus on the agricultural domain. Building domain specific knowledge graphs can be challenging due to several factors such as domain specific vocabulary, data integration challenges, dynamic data, and the need for domain expertise. Our approach primarily focuses on triplet extraction for the creation of knowledge graph. We employ dependency parsing techniques to extract relationships between entities, and utilize an extended version of BERT, combined with Latent Dirichlet Allocation (LDA) for Named Entity Recognition (NER). The proposed Agriculture knowledge graph covers significant areas of the agricultural domain by focusing on six major entities: soil, place, disease, pathogen, pesticide, and crops, along with their intra and inter-relationships. There is no benchmark dataset in the agriculture domain encompassing all the major entities. Hence we create our own corpus comprises 30k sentences sourced from reputable agriculture websites. To evaluate the effectiveness of our triplet extraction model, we utilized a test corpus containing 3500 agriculture triplets. Based on the experimental results, we were able to achieve an average macro F-score of 87% for relation extraction, indicating the efficacy of our approach. Additionally, we created an Agriculture knowledge graph using a triplet corpus of 6236 triplets. We also analyzed various knowledge graph reasoning models that improve the discovery of implicit knowledge that is not explicitly represented in the knowledge graph. Experimental results indicate that our approach is effective in creating triplets and reasoning knowledge graphs for the agricultural domain.

INDEX TERMS Agriculture, graph representation learning, knowledge graph, knowledge graph embedding, knowledge graph reasoning, named entity recognition, relation extraction.

I. INTRODUCTION

The agriculture and allied sectors are considered to be of immense importance for any economy. This is especially true for the Indian economy, where the agriculture sector provides a source of livelihood for nearly two-thirds of the population [1]. The agriculture sector involves a variety of activities, including crop cultivation, animal husbandry, forestry, and fishing, which contribute significantly to the country's economic growth [2], [3]. The climatic diversity of India enables the cultivation of a wide range of crops, including wheat, rice, maize, sugarcane, cotton, and many others. The country's topography, soil type, and rainfall patterns further contribute to the cultivation of these crops [4], [5]. Government agriculture websites provide valuable information for individuals interested in farming. These websites provide a diverse range of useful information, including the latest updates and innovations in the field.

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

These information can be extracted using the judicious use of Information Extraction (IE) tools [6]. By using these tools, farmers can quickly access the needed information, eliminating the need for hours spent manually scouring websites or research documents. These tools can filter out irrelevant or outdated information, ensuring that farmers receive only the most up-to-date and relevant information. To extract information effectively, IE tools perform subtasks such as Named Entity Recognition (NER) [7], and Relation Extraction (RE) [8]. NER identifies and classifies named entities like crops, pests, pesticides, pathogen, diseases, etc., and RE systems extracts relationships between entities. By employing these subtasks, IE tools enable the retrieval of valuable insights from agricultural sources, aiding in the provision of relevant information to farmers' queries. IE systems and Knowledge Graphs (KGs) are interconnected because the former is created to extract useful information from unstructured data while the latter stores this information in an organized and easily accessible manner.

In our study, the main focus is on automatically extracting triplets from agricultural text documents and constructing an agricultural knowledge graph. Knowledge graph representation, a crucial sub-field of Artificial Intelligence (AI) and data science, plays a vital role in expressing the structural relationships between entities in the form of facts, allowing intelligent machines to learn from knowledge graphs with ease. Knowledge graphs are a type of semantic network that represents knowledge as a collection of interconnected nodes and edges. Nodes within a knowledge graph represent real-world objects or abstract concepts, such as people, places, organizations, diseases, pesticides, soils, crops or genes. Each node is typically associated with a label and one or more attributes that describe its properties or characteristics. Edges within a knowledge graph represent the relationships between nodes. These relationships can be of different types, such as ''is a'', ''part of'', ''located in'', ''causes'', or ''suitable''. Nodes can have multiple relationships with other nodes, allowing for a rich and complex representation of the relationships within a domain. The labels associated with nodes and edges are used to capture the semantics of the relationships between objects or concepts [9]. For example, the label ''suitable'' might be used to represent the relationship between a crop and soil entity, while the label ''cause'' might be used to represent the relationship between a pathogen and a disease. Knowledge graphs [10], [11], [12] have become valuable resources for various downstream applications, viz., semantic search, personalized recommendation systems, drug discovery, chatbots and search engines. There exist generic knowledge graphs, such as BabelNet [13], YAGO [14], [15], Freebase [16], DBpedia [17], etc. to store complex structured information about the facts of the real world.

Domain specific knowledge graphs [18], [19] can be used for the purpose of arranging, handling, and exploiting vast quantities of domain specif data. Cyber [20], Biomedical [21], [22], Intelligent manufacturing [23] and Geoscience [24] are the prominent research areas of domain-specific knowledge graphs. Compared with the large amounts of online data available in agriculture, agriculture knowledge graphs are still limited. The application of knowledge graphs in the agriculture domain is expected to be a leading research direction [25]. Large language models are capable of processing and understanding natural language text to a significant extent. However, they may not be sufficient to create a comprehensive and accurate knowledge graph [26]. Creating a knowledge graph requires more than just understanding the text. It involves extracting relevant information, identifying relationships between entities, and organizing the data in a structured format. Additionally, creating a knowledge graph often requires domain-specific knowledge, which may not be readily available to the model.

Our proposed method for creating knowledge graph in the agriculture domain involves a novel triplets identification approach that can identify entities and relations from text data. Due to the absence of benchmark labelled dataset in the agriculture domain, we cannot apply supervised models in NER. Hence, we decided to investigate an unsupervised NER approach. To identify the major entities in the agriculture domain, our initial work [27], propose an unsupervised NER using Latent Dirichlet Allocation (LDA) coupled with BERT model. The algorithm is employed to recognize six major entities *Crop, Location, Soil, Disease, Pathogen, and Pesticide*. The model discovers the hidden features present in raw text automatically using the LDA topic modeling approach. The extended BERT with the LDA model creates semantically rich domain specific word vectors, which alleviates the semantic ambiguity of word vectors.

The primary emphasis of our current study lies in exploring the relationships between entities and creating a knowledge graph in the agriculture domain. Since the created knowledge graphs are incomplete in representing a huge amount of real-world facts, to further expand knowledge graphs, many types of research have been devoted to automated fact exploration. The latest advancements in knowledge graph based research concentrate on Knowledge Representation Learning (KRL) [28]. In KRL, entities and relations in a knowledge graph are represented in a low-dimensional semantic space. In our experiments, we evaluated the performance of different KRL models concerning their ability to handle Knowledge Graph Completion (KGC) [29]. Additionally, we examined their capabilities in triple classification, entity recognition, and relation extraction tasks.

The primary contributions in our work include:

- We built our proposed model on top of an existing Open Information Extraction (OIE) system that retrieves triplets from the input sentence. Once the triplets are extracted, we apply heuristic rules to verify the validity of the arguments. Our NER model labels the arguments of triplets with domain-specific tags, while the relation classifier assigns domain-specific labels to the relation phrases within those triplets.

- We utilize an unsupervised weighted distributional semantics approach for entity labeling in the agricultural domain using an extended BERT with LDA model (*exBERT _LDA+*). The combination of LDA and BERT allows us to capitalize on the respective strengths of each model. The model identifies key agricultural entities such as diseases, soil, pesticides, pathogens, crop and place.
- A relation classifier is created using a bootstrapping technique that relies on identifying the shortest path between entities within an undirected dependency graph. This study mainly focuses on six inter subdomain relations such as <*Soil,Location*>, <*Soil,Crop*>, <*Disease,Crop*>, <*Disease,Pathogen*>, <*Pathogen, Crop*>, <*Pesticide,Pathogen*> and four intra subdomain relations such as <*Disease,Disease*>, <*Pathogen, Pathogen*>, <*Soil,Soil*>, <*Pesticide,Pesticide*> that link entities Crop, Place, Disease, Pathogen, Soil, and Pesticide in the agriculture domain.
- We propose an automatic knowledge graph creation in the agriculture domain based on the triplets(Entity1, Rel, Entity2) created using entity and relation classification models. We utilise the Knowledge Graph Embedding models to handle one-hop, multi-hop and missing binary relations present in the agriculture knowledge graph.
- The absence of a benchmark dataset motivated us to create an agriculture corpus of 30000 sentences for NER, which can be used for other NLP tasks. Our approach identified approximately 6,236 triplets from the agricultural sentence corpus, which is utilized by KGE models for knowledge graph creation and reasoning.

The paper is arranged as follows. The review of the related works is presented in Section II. In Section III, we introduce the proposed method and outcomes of our experiments. The limitations of the proposed model is discussed in Section IV. In Section V, we summarize the main findings of our study and explore further research in the field.

II. RELATED WORKS

In order to construct a comprehensive information extraction system within the agriculture domain, we conducted an in-depth literature survey focusing on existing approaches for Named Entity Recognition, Relation Extraction, Knowledge Graph Creation and Completion. The following sections discuss the findings in detail.

A. NAMED ENTITY RECOGNITION

The concept of NER was first introduced during the sixth Message Understanding Conference [30]. Since then, various online tools available for recognizing open domain named entities. Some examples of such tools are StanfordCoreNLP [31], Natural Language ToolKit (nltk) [32], OpenNLP [33], IBM Watson [34], NeuroNER [35], displaCy [36], AllenNLP [37], and AmazonComprehend [38]. When it comes to domain-specific applications such as biomedical, finance, cyber, and agriculture the performance of these models tends to be less accurate. The rule-based NER systems [39], [40], [41] rely on hand-crafted rules to recognize named entities. The process of manually constructing rules demands a considerable amount of time and effort. Statistical based NER systems [42], [43], [44] utilize a substantial amount of labeled data to train the NER model. Recently, the deep learning systems [45], [46], [47], [48], etc have achieved state-of-the-art results compared to traditional machine learning models. However, effectively adapting these NER models to agriculture NER applications remains a substantial challenge. After conducting a comprehensive literature survey, we discovered that most existing approaches rely on supervised setups, which require vast amounts of data.

B. RELATION EXTRACTION

Extensive research is conducted on supervised approaches to address domain specific relation extraction such as [49], [50], [51], and [52]. Asma and Pierre [53] propose an approach to extract relations between diseases and treatments from MEDLINE data. Sharma et al. [54] present a verb-focused approach for extracting relationships from biomedical text. Vani et al. [55] propose a novel method that employs the Shortest Dependency Path (SDP) to choose the most representative samples. Additionally [56] present RelEx a Dependency parsing-based relation extraction system in the Biomedical domain. There are other trained RE systems viz., [48], [50], [57], [58] in the Biomedical domain. The performance of such supervised approaches heavily relies on the availability of labeled data. Accurately recognizing domain-specific relations with existing models is challenging due to the semantic complexity associated with the relation phrases. Hence the adaptability of domain specific trained models to the agriculture domain poses a significant challenge.

Another direction of relation extraction study belongs to semi-supervised learning methods [59], [60]. The line of research begins with DIPRE, [61], which specifically designed to extract $\langle book, author \rangle$ relations. A similar methodology of DIPRE is proposed in Snowball [62]. A combination of DIPRE with distributional semantics is proposed in BREDS by Batista et al. [63]. BREDS incorporates ReVerb System [64] to identify the relation pattern. In order to extract relations at a larger scale, distant supervision methods are also utilized [65]. The research works [66] and [67] provide an extensive survey on relation extraction techniques using distant supervision.

The Open Information Extraction(OIE) Systems which utilize linguistic aspects of the input data to extract relational triplets from a given corpus [68] are utilized to solve the challenges of labeled samples. Different types of such systems are available viz., Learning based [69], Rule based [70], and Clause based [71]. Dependency parsing based is utilized for relation extraction in Stanford Open IE [72]. The OIE Systems which focus on linguistic knowledge of

input corpus are ClausIE [71] and MinIE [73]. An in depth examination of OIE Systems can be found in [74] and [75]. These systems are capable of extracting relations and arguments in the form of triplets from a given sentence. However, they are unable to label the arguments and relation present in the triplets using domain knowledge.

C. KNOWLEDGE GRAPH CREATION AND COMPLETION

The approaches for automatic construction of knowledge graphs and completion from data sources are reviewed in this section. An application that performs automatic generation of custom knowledge graph using dynamic extraction of triplets and ontology is presented in [76]. The created knowledge graph are used for applications domains such as customer support, HR, banking, journalism. A hybrid approach that combines a rule-based approach and a similarity-based approach is proposed in T2KG [77]. A semiautomatic method that generates ontology-linked knowledge graph from biomedical texts is proposed in [78]. An end-to-end system is proposed in [79] to automatically extract DBpedia RDF triples from Wikipedia text using SRL and coreference resolution techniques. An in-depth examination of knowledge graph creation can be found in the papers [80], [81], and [82], etc. We review literature that examines Knowledge Graph Completion(KGC) techniques aimed at expanding existing knowledge graphs. To enhance existing knowledge graphs, there are Knowledge Graph Embedding (KGE) models which predict missing relations between the entities in knowledge graphs. KGE models, which learn semantic representations of entities and connections, are usually used to solve the missing link (relation) prediction or KGC problems. KGE models have been categorised into three groups such as translational distance based, semantic matching based, and neural network based. Given a triple (Entity1, Rel, Entity2), translational distance based models translates the head entity *Entity1* to the tail entity *Entity2* using the relation *Rel*. TransE [83] is the simple and effective translational distance model achieving the state-of-the-art link prediction performance. But the model failed to model 1-to-N, N to-1 and N-to-N relations. To overcome the disadvantages of TransE, TransH [84] introduced relation-specific hyperplanes. TransR [85] shares a similar concept to TransH. It introduces relation-specific spaces and conducts the translation within the relation space. For each relation *Rel*, it uses matrix M_{rel} to map entity embeddings to the relation space. Compared to TransE and TransH, TransR requires a large amount of computing resources. Since these models failed to recognize composition relations, RotatE [86] is proposed to identify such relations. To capture the semantic hierarchies HakE is proposed [87]. The second group of models, semantic matching models use similarity based scoring function to calculate the similarity between the different entities and relations. The translational based multiplicative models which capture more semantic information are DistMult [88], HoIE [89] and Complex [90]. The third category of KGE Models use Convolutional Neural Network (CNN) framework for knowledge graph completion [91], [92]. Majority of existing KGE models failed to recognize all three important relation patterns, symmetry/antisymmetry, inverse and composition. There are models proposed to identify such relations RotatE [86], MuRE [93], Tucker [94], and PairRE [95]. The papers [10], [96], [97], and [98] provide a comprehensive analysis.

D. AGRICULTURE SPECIFIC SYSTEMS

In the agriculture domain, there is a scarcity of tools and models specifically developed for entity tagging and relation extraction tasks viz., [99], [100], [101]. Hongchen and Yiheng [25] proposed an approach that utilizes BERT+ BiGRU + CRF to extract the entity and relationship present in an input sentence. In our initial work [102] we propose LDA based NER model for recognizing agriculture entities. We also introduced a bootstrapping approach for relation extraction [103], which identified five inter-subdomain relations. Guo et al. [104] propose a supervised method for agricultural NER utilizing BERT-based contextual embeddings and glyph features extracted by a 3D Convolutional Neural Network. By applying this approach to the self-annotated corpus known as AgCNER, the diseases and pests dataset in the Chinese agricultural domain, an F1 score of 95% is achieved. Du et al. [105] proposed Soil ontology focusing soil characteristics and processes. The Government of India has undertaken several initiatives to respond to inquiries from farmer's in their native language, such as *eSagu* [106], *aAQUA* [107], and the Kisan Call Centres (KCCs)[1,2].

E. SUMMARY AND RESEARCH GAPS

After an extensive literature review, it was noted that although there are numerous agriculture-related repositories, there is currently a lack of a standardized benchmark data set specifically designed for the agriculture domain. Hence, it becomes essential to explore potential information extraction tools that can effectively extract pertinent and valuable information from the extensive repositories of agricultural data. Based on our analysis, it was found that there is a lack of standardized methodologies for creating an Agriculture Knowledge Graph(AgriKG) from text data. Additionally, many of the existing domain-specific knowledge graphs are not publicly available, making it difficult for other researchers to access and use them.

Considering these identified gaps in the literature, the focus of the current work is centered around the development of a knowledge graph in the agriculture domain. To perform entity identification we use OIE system, while extended BERT with LDA model is used for NER. Our approach introduces semi-supervised bootstrapping that leverages the dependency parse tree to identify relation patterns. We also construct a corpus

[1] https://mkisan.gov.in/aboutkcc.aspx
[2] manage.gov.in/kcc/kcc.asp

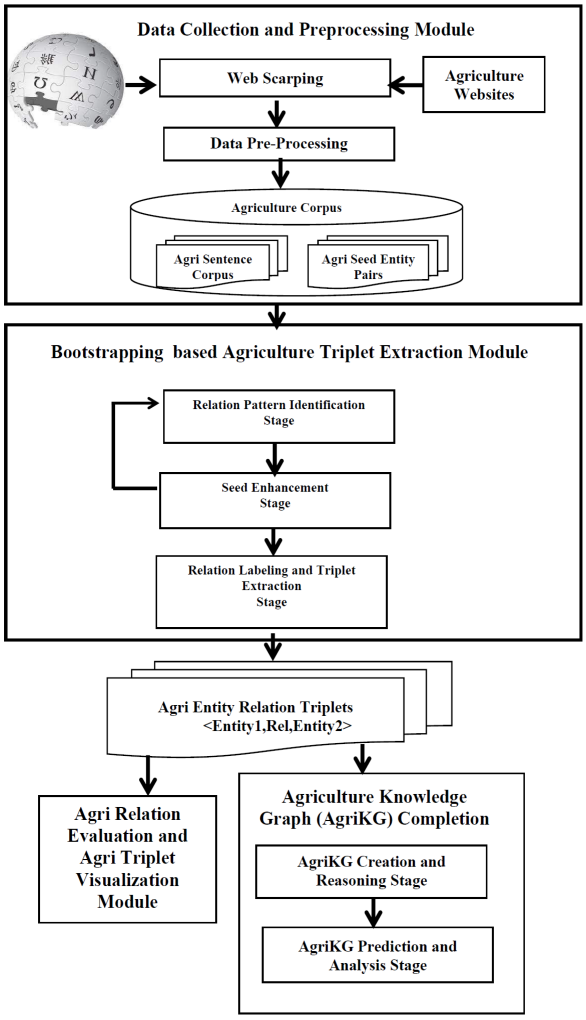


FIGURE 1. The Pipeline model of the proposed automatic agriculture knowledge graph creation and reasoning system.

specifically related to agriculture to facilitate the proposed evaluations. The methodology used in this study, outlining the step-by-step process of the pipeline model is presented in the following Section III.

III. METHODOLOGY

The proposed Automatic Agriculture Knowledge Graph Creation and Completion System presented in Figure 1, is designed to extract 10 inter/intra relations in the agriculture domain. The entire process is divided into four major modules, viz., Data collection and preprocessing module, Bootstrapping based Agriculture Triplet Extraction module, Agriculture relation evaluation and triplet visualization Module, and Agriculture Knowledge Graph Completion module. In module 1, a dataset exclusively focused on agriculture was created by extracting sentences from reputable agricultural websites and Wikipedia. This agriculture corpus is utilized in module 2, which follows a semi-supervised bootstrapping approach for extracting triplets. In module 3, we assess the relationships and visualize the triplets. These triplets are utilized in module 4 to facilitate AgriKG creation, reasoning and evaluation. The detailed algorithm for agriculture knowledge graph creation can be found in Algorithm 1. Algorithm 1 incorporates a submodule known as Globalvector(), and this specific submodule is presented in a separate Algorithm 2. Also, Algorithm 2 includes calls to both Algorithm 3 and Algorithm 4. The subsequent subsections provide a detailed discussion of the entire pipeline depicted in Figure 1.

Algorithm 1 AgriKG(AgriSentCorpus[1..n], SeedEntityPairs[1..m])

Input: *SeedEntityPairs* represent seed entity pairs in all subdomains, *AgriSentCorpus* is the agriculture sentence corpus
Output: Knowledge Graph

$GRV_{Subdomian} \leftarrow GlobalVector(AgriSentCorpus, SeedEntityPairs)$
while $i < n$ **do**
  y ← AgriSentCorpus[i]
  Generate Triplets of the sentence *y*,<Entity1, Rel, Entity2> using OIE system
  l1 ← AGRONER(Entity1) using Equation (3) l2 ← AGRONER(Entity2) using Equation (3)
  Create Local Relation vector $LRV_{Subdomian}$ for *Rel* using Equation (4)
  Calculate the cosine similarity **score** of $LRV_{Subdomian}$ with $GRV_{Subdomian}$ using Equation (5)
  /* $\theta_{\text{subdomain}}$ is the average similarity score of all relation patterns present in an inter subdomain with the GRV */ **if** *score* $> \theta_{subdomain}$ **then**
    RelLabel ← RLabel(l1,Score,L2), using Equation (6)
    Add<Enity1, RelLabel, Entity2> to Agri triplet corpus
  end
  Create Agriculture Knowledge Graph (AgriKG) using Agri triplet corpus
  return AgriKG
end

A. DATA COLLECTION AND PRE-PROCESSING MODULE

In the agriculture domain, there is no comprehensive benchmark dataset available that covers all the prominent entities. However, there are agriculture datasets available, which serve specific purposes. We analyzed the various datasets, it was found that the majority of the agricultural data consist of images and numerical or sensor dataset values. Hence, a dataset exclusively focused on agriculture domain was created by scraping agricultural related sentences from Wikipedia and reputable agricultural websites. To accomplish this, the first stage of the data collection process, as shown in Figure 1, involved extracting information from authorized

TABLE 1. Data statistics of the agri seed entity pairs in the inter-intra subdomains.

Sl:No	Inter-Intra Subdomain	No of Seed Entities
1	<Soil,Crop>	25
2	<Soil,Location>	25
3	<Pathogen,Crop>	150
4	<Disease,Crop>	150
5	<Disease, Pathogen>	1000
6	<Pesticide, Pathogen>	100
7	<Soil,Soil>	20
8	<Pesticide, Pesticide>	100
9	<Pathogen,Pathogen>	1000
10	<Disease,Disease>	2100

websites using seed pairs related to subdomains such as Soil, Disease, Pathogen, and Pesticides. For example, key phrases related to plant diseases and pathogens were gathered from *The American Phytopathological Society* [108] website. To ensure data quality and consistency, basic pre-processing steps were performed using the Natural Language Toolkit (NLTK). The Python Beautiful Soup Algorithm is used to scrape the relevant information from websites. This agri sentence corpus comprises 30,000 sentences. Further insights into the process of Data collection and preprocessing are available in our previous research [27]. The collected agri sentence corpus is utilized in the Agriculture Triplet Extraction Module. We manually collected some Seed Entity Pairs which is also required for triplet extraction. Data statistics of the Seed Entity Pairs in the inter-intra subdomains are shown in Table 1. As an example, consider the entity pair ⟨ Angular leaf spot, Strawberry ⟩ represents a seed entity pair within the inter subdomain ⟨*Disease*, *Crop*⟩. The relationship between the disease name and the corresponding affected crop is represented by this entity pair. Likewise, ten pairs of seed entities are selected for inter-intra subdomains using a similar approach.

B. BOOTSTRAPPING BASED AGRICULTURE TRIPLET EXTRACTION MODULE

This module describes the bootstrapping approach for extracting triplets from the agri sentence corpus by using the agri seed entity pair. The entire process is divided into three stages viz., Relation Pattern Identification(Stage 1), Seed Enhancement(Stage 2) and Relation Labeling and Triplet Extraction(Stage 3). Following subsections illustrate the triplet extraction process in detail.

1) RELATION PATTERN IDENTIFICATION STAGE

This module focuses on extracting the relationship patterns present in a sentence involving seed entities. Figure 2 illustrates the pipeline model employed, showcasing the various stages involved in identifying and analyzing the relation patterns. The detailed procedure for relation pattern identification stage is presented in Algorithm 3. As shown in Figure 2, there are two inputs to this module named *Agriculture corpus* and *Seed entity pairs*. In our study, we took six subdomains in the agriculture domain, viz,. *crop, location, disease, soil, pathogen, and pesticide*. Also, six main inter-domain relations viz,. ⟨*Soil*, *Crop*⟩, ⟨*Soil*, *Location*⟩, ⟨*Pathogen*, *Crop*⟩, ⟨*Disease*, *Pathogen*⟩, ⟨*Disease*, *Crop*⟩, and ⟨*Pesticide*, *Pathogen*⟩.

Algorithm 2 GlobalVector(AgriSentCorpus[1..n], SeedEntityPairs[1..m])

Input: *SeedEntityPairs* represents seed entity pairs in all subdomains and *AgriSentCorpus* is the agriculture sentence corpus
Output: Global Relation Vector
Relation$_{Subdomian}$ ←{ } /*To store the list of Relation Patterns present in each subdomains;*/
while *i*<*m* **do**
  < *Entity*1, *Entity*2 > ←SeedEntityPairs[i];
  while *j*<*n* **do**
    if *Sentence* S_j *from the Agri Sentence Corpus contains both Entity1 and Entity2* **then**
      Identify the relation pattern within the sentence S_j
      RelPhrase ← **RelPatternIdentification**(S_j, Entity1, Entity2)
      Add RelPhrase to *Relation*$_{Subdomian}$
      L1 ← EntityLabeling(Entity1)
      L2 ← EntityLabeling(Entity2)
      seedpair ← **SeedEnhancement**(L1, Relphrase, L2)
    end
    Add seedpair to SeedEntityPairs list
  end
end
Create Global Relation vector *GRV*$_{Subdomian}$ for *Relation*$_{Subdomian}$ using Equation (2);
return *GRV*$_{Subdomian}$

Examples of inter-intra subdomains, seed entity pairs, and sample sentences are shown in Table 2. During the initial stage of the bootstrapping process, sentences with entity pairs are identified from the agri sentence corpus. The candidate sentences are displayed in the fourth column of Table 2. These sentences are carefully chosen as they contain information pertinent to the specific entities under consideration. The seed pairs in the extracted sentences are then labeled into six categories, viz.,*crop, place, disease, soil, pathogen and pesticide*. Since we use the seed entities for sentence extraction, we are sure about the named entities. Hence at this stage, we are not applying any models for NER.

Once the named entities are labeled, the sentence undergoes dependency parsing. Dependency parsing, as described in [109] identifies the semantic relationships between words within a sentence. As an example, consider the given sentence *"Laterite soil, which is rich in calcium and potash, is suitable for growing cotton"*, which is extracted using the seed entity

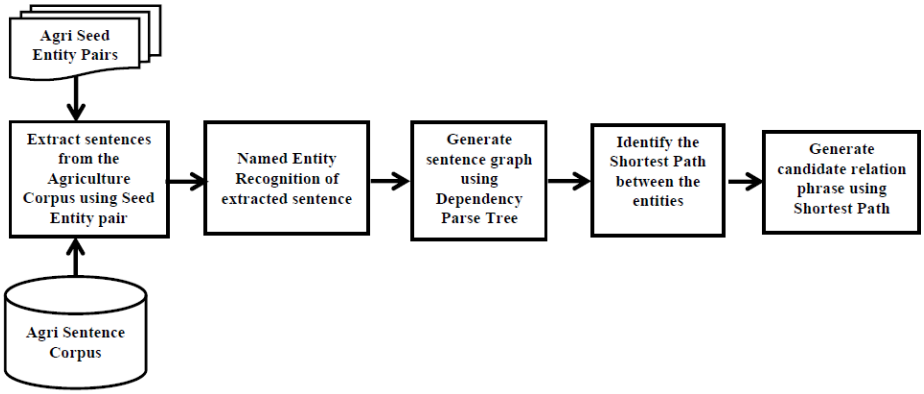
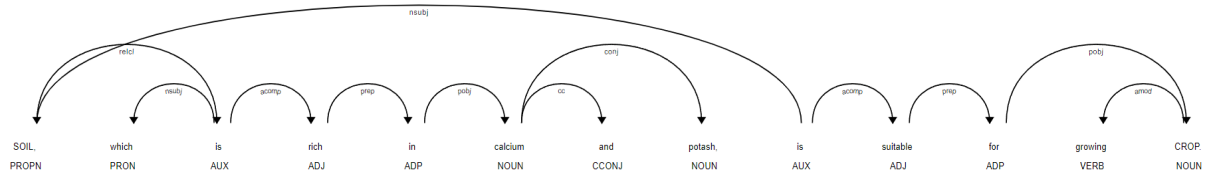


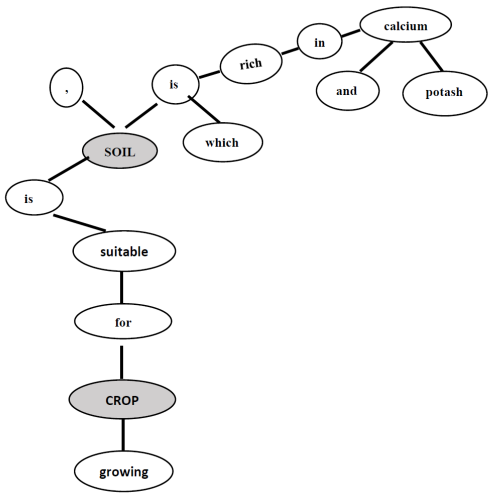
FIGURE 2. The pipeline model of the relation pattern identification stage.



[A]

Sl:No	Token	Head	Dependents
1	SOIL	is	is
2	which	is	
3	is	SOIL	rich,which
4	rich	is	in
5	in	rich	calcium
6	calcium	in	and,potash
7	and	calcium	
8	potash	calcium	
9	is	SOIL	suitable
10	suitable	is	for
11	for	suitable	growing
12	growing	CROP	
13	CROP	for	growing

[B]

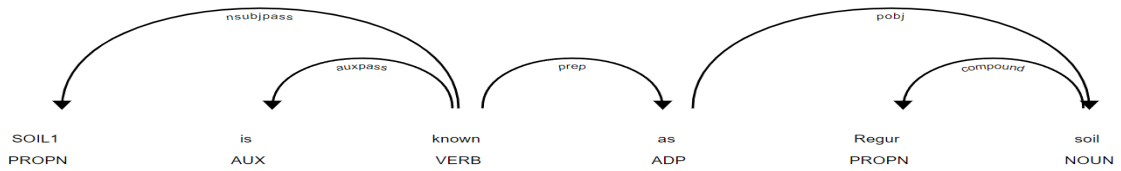


[C]

FIGURE 3. [A]: Dependency tree (inter subdomain) of the sentence *"SOIL, which is rich in calcium and potash, is suitable for growing CROP"*. [B]: Tokens with Head and Dependent for the same sentence. [C]: Dependency Graph of the same sentence.

pair <Laterite soil, Cotton>. The given entity pair indicates the type of soil that is appropriate for a particular crop. After NER labeling of the seed entities, the sentence is changed to *"SOIL, which is rich in calcium and potash, is suitable for growing CROP"*. The dependency parse tree of the sentence *"SOIL, which is rich in calcium and potash, is suitable for growing CROP"* is presented in Figure 3 A. The head-dependent relation found in the dependency tree is utilized to generate an undirected dependency graph. In Figure 3 B, tokens along with their associated head-dependent words are presented. Then using the head-dependent relations, an undirected dependency graph is created. The dependency graph is shown in Figure 3 C. Consider another example from the intra subdomain <Soil, Soil>, the input sentence is ''Black soil is also known as regur soil''. After NER labeling the sentence becomes *"SOIL1 is also known as regur soil"*. Entity1 is labeled as SOIL1. Dependency parse tree presented in Figure 4 A. In Figure 4 B, tokens along with their associated head-dependent words are presented. In 4 C undirected dependency graph is presented. We use Spacy[3] [36] for dependency parsing.

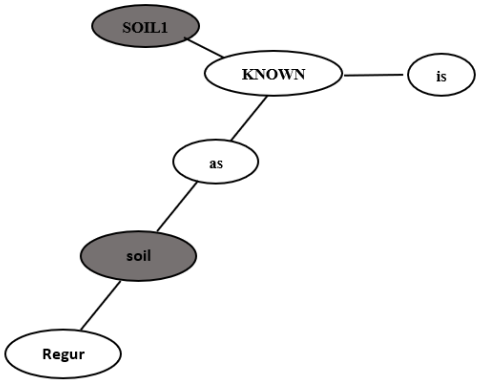
[3]https://spacy.io/



[A]

Sl:No	Token	Head	Dependents
1	SOIL1	known	
2	is	known	
3	known		SOIL1,is,as
4	as	known	soil
5	Regur	soil	
6	soil	as	Regur

[B]



[C]

FIGURE 4. **[A]: Dependency tree (intra subdomain) of the sentence *"SOIL1, is known as regur soil"* [B]:Tokens with Head and Dependent of the same sentence. [C]: Dependency Graph of the sentence.**

Algorithm 3 RelPatternIdentification(S,Entity1, Entity2)

Input: Sentence S with n words($w_1, w_2, \ldots..w_n$), <Entity1, Entity2> represents the seed pair
Output: Relation Phrase *RelPhrase* present in the sentence S

29 $D_{tree} \leftarrow \{\}$
/* The DependencParse Tree associated with the Sentence S1*/
n is the set of tokens present in the in *D_tree*
E is the set of edges present in the Dependency tree *D_tree*
30 Create S1 by eliminating any special characters and stop words present in the input sentence S
31 Generate Dependency Parse Tree *D_tree* for the Sentence S1
32 **while** $i < |S1|$ **do**
33 Identify the word *W_head* associated with the word token w_i
34 Create an edge *e* using *W_head* and w_i
35 Add *e* to the list of Edges *E*
36 **end**
37 Create a graph *G* using the Edges *E*
38 **RelPhrase**=BreadthFirst Search(G,Entity1,Entity2)
39 return Relphrase

Moving on to the next step, the words between seed entities are identified. We utilize the Shortest Dependency Path (SDP) to extract the words between the seed entities and this path is determined using Breadth First Search(BFS) of dependency graph. In the given example BFS algorithm is utilized to traverse the graph using the search pattern (CROP, SOIL) and the potential relation between the entities is identified as **is suitable for**. In cases where a path between the entities cannot be established, all the words located between them are extracted for further analysis. For example, consider the sentence *"Brown rot is a common and destructive disease affecting apricot"*. The model output *"is a common and destructive disease affecting"* as the relation phrase. In Table 3, the relation patterns extracted during the relation pattern identification stage is presented. In the following subsection, the explanation is provided on how the extraction of additional entity pairs can be facilitated through the utilization of these relation phrases.

2) SEED ENHANCEMENT STAGE

In this module, the relation patterns extracted in the previous step are utilized to enhance the seed entities. Using these relation patterns along with entity labels, additional sentences are extracted from the agriculture corpus. These newly acquired sentences contribute to expand the seed entities and further refining their context and associations within the agricultural domain. The process is explained in detail in Figure 5 and in Algorithm 4.

The extracted new sentences are passed on to an OIE System viz., MinIE [68], [71], [73], which identifies the triplets ($< Entity1, Rel, Entity2 >$) that are present within sentence. The OIE system extracts triplets that consist of the arguments and their corresponding relations. These arguments represent the named entities in the input sentence.

TABLE 2. Seed entities and candidate sentences present in inter-intra subdomains.

Sl:No	Inter-Intra Subdomain	Seed Entities	Candidate sentences
1	$\langle Soil, Crop \rangle$	$\langle LateriteSoil, Cotton \rangle$	Laterite soil, which is rich in calcium and potash,is suitable for growing cotton.
2	$\langle Soil, Location \rangle$	$\langle BlackSoil, Madurai \rangle$	Black soils are found extensively in the districts of Madurai
3	$\langle Pathogen, Crop \rangle$	$\langle Papayaringspotvirus, Papaya \rangle$	Papaya ringspot virus causes a major disease of papaya.
4	$\langle Disease, Crop \rangle$	$\langle Angularleafspot, Strawberry \rangle$	Angular leaf spot is frequently problematic in strawberry.
5	$\langle Disease, Pathogen \rangle$	$\langle Angularleafspot, Xanthomonas \rangle$	Angular leaf spot is a bacterial disease caused by Xanthomonas fragariae.
6	$\langle Pesticide, Pathogen \rangle$	$\langle Gliotoxin, Rhizoctoniasolani \rangle$	Gliotoxin is an antagonistic effect toward R. solani
7	$\langle Soil, Soil \rangle$	$\langle BlackSoil, Soil \rangle$	Black soil is known as Regur soil.
8	$\langle Pesticide, Pesticide \rangle$	$\langle Dinoseb, Herbicide \rangle$	Dinoseb is a herbicide that was once widely used.
9	$\langle Pathogen, Pathogen \rangle$	$\langle Papayaringspotvirus, Virus \rangle$	Papaya ringspot virus is a pathogenic plant virus.
10	$\langle Disease, Disease \rangle$	$\langle Anthracnose, Disease \rangle$	Anthracnose, a group of fungal diseases that affect a variety of plants in warm, humid areas.

TABLE 3. Candidate sentences and corresponding relation patterns.

Sl:No	Candidate sentences	Relation Pattern
1	Laterite soil, which is rich in calcium and potash,is suitable for growing cotton.	is suitable for
2	Black soils are found extensively in the districts of Madurai	found in districts of
3	Papaya ringspot virus causes a major disease of papaya.	causes disease of
4	Angular leaf spot is frequently problematic in strawberry.	is problematic in
5	Angular leaf spot is a bacterial disease caused by Xanthomonas fragariae.	is disease caused by
6	Gliotoxin is an antagonistic effect toward R. solani	is effect toward
7	Black soil is known as Regur soil	known as
8	Dinoseb is a herbicide that was once widely used	is
9	Papaya ringspot virus is a pathogenic plant virus.	is
10	Anthracnose, a group of fungal diseases that affect a variety of plants in warm, humid areas.	group of

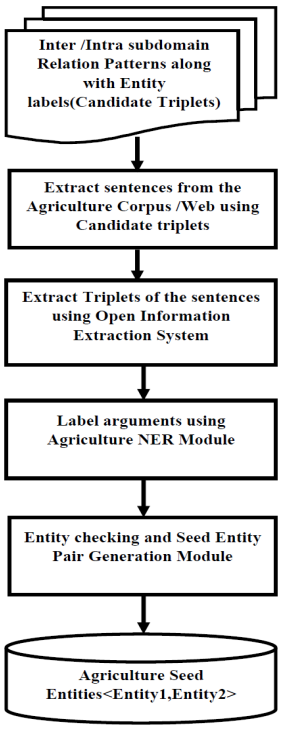


FIGURE 5. Pipeline model of seed enhancement stage.

After the extraction of triplets, we employ heuristic rules to validate the extracted arguments. These rules filter out entities that are excessively long, begin with a preposition, or are linking verbs. The arguments/entities within the triplet are labeled using our prior work on NER named AGRONER [27].

As an example, the relation pattern *"is suitable for"* is extracted from the sentence *"Laterite soil, which is rich in calcium and potash, is suitable for growing cotton."* as shown in Table 3. Using this relation pattern and entity labels a new search pattern is identified as <SOIL, is suitable for, CROP>. Consequently, the sentence *"Loamy soil is ideal for growing crops such as wheat, sugarcane, cotton, jute, pulses, and oilseeds"* is extracted from agri sentence corpus. In the

Algorithm 4 SeedEnhancement(L1, RelPhrase,L2)

Input: L1 and L2 represent entity labels, RelPhrase represent the relation pattern
Output: Seed Entity Pairs present in the Agri Sentence Corpus

Extract Sentence S using the search pattern <L1, RelPhrase, L2> from Agri sentence corpus or Web
Create S1 by eliminating any special characters and stop words present in the input sentence S
Generate Triplets <Entity1, Relation, Entity2> of S1 using OIE system;
l1 ← AGRONER(Entity1)
l2 ← AGRONER(Entity2)
if l1 and l2 ∈ {S, L, PA, PE, C, and D}
    add l1 and l2 to *SeedPair*
return *SeedPair*

TABLE 4. Triplets Extracted using OIE system for the sentence *"Loamy soil is ideal for growing crops such as wheat, sugarcane, cotton, jute, pulses, and oilseeds."*

Sl:No	Entity1	Rel	Entity2
1	Loamy soil	is ideal for growing crops such as	wheat
2	Loamy soil	is ideal for growing crops such as	sugarcane
3	Loamy soil	is ideal for growing crops such as	cotton
4	Loamy soil	is ideal for growing crops such as	jute
5	Loamy soil	is ideal for growing crops such as	pulses
6	Loamy soil	is ideal for growing crops such as	oilseeds
7	wheat	is	crop
8	sugarcane	is	crop
9	cotton	is	crop
11	jute	is	crop
12	pulses	is	crop

absence of relevant sentences in the corpus, we use Scrapy[4] to extract sentences from the web. The triplets of these new sentences are generated using OIE system and the results are presented in Table 4. The arguments/entities are labeled using AGRONER.

In AGRONER, an unsupervised NER using weighted distributional semantic model is proposed to identify the major entities in the agriculture domain. In AGRONER, we make use of the AGROVOC dictionary lookup, as described in [110] and [111], to recognize crops. Since place entities are closely linked to crops, it is crucial to identify them for our research. To accomplish this, we employ the geocoding web service called *geopy* in Python. The remaining entities *Soil,Disease, Pathogen, Pesticide* are identified using LDA coupled with BERT model. BERT offers several advantages over Word2Vec and GloVe for natural language processing tasks. In contrast to Word2Vec and GloVe, BERT is pretrained on extensive text corpora, emerging as a robust language model with leading-edge performance. BERT reduces ambiguity by leveraging a bidirectional approach during pretraining, considering both left and right context to develop a contextualized understanding of words. As a result, BERT excels in providing contextual understanding of words and sentences, allowing it to capture nuances in meaning across different contexts and handle polysemy effectively. As our dataset lacks labels, we utilize topic modeling to uncover

[4]https://docs.scrapy.org/en/latest/

TABLE 5. Subdomains and relation patterns.

Sl:No	Inter/Intra Subdomains GlobalVector	Relation Patterns
1	$\langle Soil, Crop \rangle$	is required for growing are ideal for the growth of in thrive carrots is appreciate is preferred for is ideal for growing crops have problems growing
2	$\langle Soil, Location \rangle$	distributed across districts of found in districts of place is found mostly in areas of found mostly in the southern parts of covers
3	$\langle Disease, Pathogen \rangle$	caused by disease of fruits caused by disease caused by is disease affects fruit as recognized symptoms caused by caused by members of pathogen
4	$\langle Pathogen, Crop \rangle$	causes disease is pathogen causes blight on infects beset by threatens reported on
5	$\langle Pesticide, Pathogen \rangle$	is fungicide used in plantations known for control of diseases used in control used in cropping recommended
6	$\langle Disease, Crop \rangle$	attacks is disease in developing in is disease affecting is a common disease of is a serious disease of
7	$\langle Pesticide, Pesticide \rangle$	is is fungicide is member of is representative of
8	$\langle Pathogen, Pathogen \rangle$	is is a saprotrophic is a devastating bacterial is pathogen of agricultural crops was bacterium proven be is the most common
9	$\langle Disease, Disease \rangle$	is is one of the major is a common is one of the most important is a notifiable plant is a blossom-infecting fungal is one of the most destructive
10	$\langle Soil, Soil \rangle$	is is type of is kind of known as

the underlying topics within the corpus. We have used the most popular approach LDA [112]for topic modeling. Later these topics are vectorized using BERT, a semantic model that utilizes distributional properties of words. The combination of LDA and BERT allows us to capitalize on the respective strengths of each model. Since each entity class has its own unique dataset, we apply the LDA Topic Modeling module separately to each dataset. Details of LDA parameters are presented in the Appendix, Table 15. The default values provided by Gensim library is used for hyper parameters. This approach helps us to capture the specific characteristics of each entity class. The number of topics in the dataset are determined using coherence score. These topics are then vectorized using BERT to create a Global Vector. To handle domain-specific entities, the BERT tokenizer is extended using domain specific vocabulary. As our focus was on four entities, we generated four Global Vectors. For NER, the arguments of the triplet are vectorized using BERT. These local vectors are transformed into weighted vectors using scores obtained from the LDA model. The labeling process is based on the cosine similarity between these local vector and the global vectors. For instance <Loamy soil, Wheat> present in Table 4 is labeled as <SOIL,CROP> using AGRONER module. Therefore, <Loamy soil, Wheat> is considered as a potential seed pair for extracting more seed pairs and relation patterns. We have applied the similar procedure to the remaining eleven triplets as shown in the Table 4. In order to store the extracted relations patterns and seed entity pairs, distinct clusters are maintained for each inter-subdomain. The extracted relation patterns are assigned to the appropriate clusters based on their associated entity labels. Table 5 shows sample relation patterns and the corresponding subdomains. The extracted relation patterns shown in Table 5 are subsequently utilized for relation labeling and final triplet identification, as elaborated in following subsection.

3) RELATION LABELING AND TRIPLET EXTRACTION STAGE

Figure 6 illustrates the pipeline model of the Relation Labeling and Extraction Stage. All the inter/intra subdomain relation patterns identified in the bootstapping stage are gathered and transformed into vector representations as shown in Figure 6. The vector representations are generated using the BERT Base model [113]. Examples of relation phrases are provided in Table 5. Using these relation patterns, a Global Relation Vector(GRV) is created for each inter/intra subdomains. The GRV represents the global significance of a relation in a Corpus, which is computed as shown in Equation (1) and (2). The Equation (1) is used to transform a relation phrase into a vector representation. Next, the Global Relation Vector is calculated by taking the average of all subdomain relation vectors.

$$Vec(rp) = \frac{\sum_{i=1}^{N} BERT\_Vec(w_i)}{N} \tag{1}$$

$$GRV(rp) = \frac{\sum_{i=1}^{R} Vec(rp_i)}{R}, \forall rp \in Relation\_subdomain \tag{2}$$

$$= \begin{cases} GRV\_S\_L; \text{ if } \forall rp \in Relation\_subdomain\_S\_L \\ GRV\_S\_C; \text{ if } \forall rp \in Relation\_subdomain\_S\_C \\ GRV\_D\_PA; \text{ if } \forall rp \in Relation\_subdomain\_D\_PA \\ GRV\_D\_C; \text{ if } \forall rp \in Relation\_subdomain\_D\_C \\ GRV\_PA\_C; \text{ if } \forall rp \in Relation\_subdomain\_PA\_C \\ GRV\_PE\_PA; \text{ if } \forall rp \in Relation\_subdomain\_PE\_PA \\ GRV\_D\_D; \text{ if } \forall rp \in Relation\_subdomain\_D\_D \\ GRV\_PA\_PA; \text{ if } \forall rp \in Relation\_subdomain\_PA\_PA \\ GRV\_S\_S; \text{ if } \forall rp \in Relation\_subdomain\_S\_S \\ GRV\_PE\_PE; \text{ if } \forall rp \in Relation\_subdomain\_PE\_PE \end{cases}$$

$$l_e = AGRONER(entity_e) : e \in \{1, 2\} \tag{3}$$

$$LRV(input\_rp) = \frac{\sum_{i=1}^{M} BERT\_Vec(w_i)}{M} \tag{4}$$

$$Score = Cos(LRV\_i, GRV\_i) : i \in \{\text{S\_L, S\_C, D\_C, D\_PA, PA\_C, PE\_PA, S\_S, PE\_PE, PA\_PA, D\_D}\} \tag{5}$$

$$RLabel(l1, Score, l2) = \begin{cases} Suitablefor; \text{ if } l1 = S, l2 = C \text{ and } Score \geq \theta_{\text{S\_C}} \\ Found; \text{ if } l1 = S, l2 = L \text{ and } Score \geq \theta_{\text{S\_L}} \\ Causedby; \quad \text{if } l1 = D, l2 = PA \text{ and } Score \geq \theta_{\text{D\_PA}} \\ Infect; \quad \text{if } l1 = PA, l2 = C \text{ and } Score \geq \theta_{\text{PA\_C}} \\ Used; \quad \text{if } l1 = PE, l2 = PA \text{ and } Score \geq \theta_{\text{PE\_PA}} \\ Affect; \quad \text{if } l1 = D, l2 = C \text{ and } Score \geq \theta_{\text{D\_C}} \\ isaSoil; \quad \text{if } l1 = l2 = S \text{ and } Score \geq \theta_{\text{S\_S}} \\ isaPesticide; \quad \text{if } l1 = l2 = PE \text{ and } Score \geq \theta_{\text{PE\_PE}} \\ isaPathogen; \quad \text{if } l1 = l2 = PA \text{ and } Score \geq \theta_{\text{PA\_PA}} \\ isaDisease; \quad \text{if } l1 = l2 = D \text{ and } Score \geq \theta_{\text{D\_D}} \\ OTHER; \text{ Otherwise} \end{cases} \tag{6}$$

This computation is illustrated in Equation (2). In the equations, we have use the labels C, L, D, PA, PE, and S to denote the respective subdomains Crop, Location, Disease, Pathogen, Pesticide, and Soil. Let $Relation\_subdomain = \{rp_1, rp_2, \ldots, rp_R\}$, store all relation patterns of a particular subdomain, $rp = \{w_1, w_2 \ldots w_n\}$ represents the relation pattern where w_i refers to the ith word of the relation pattern, R represents the total number of relations,and N denotes the total word count of the relation pattern. For instance,

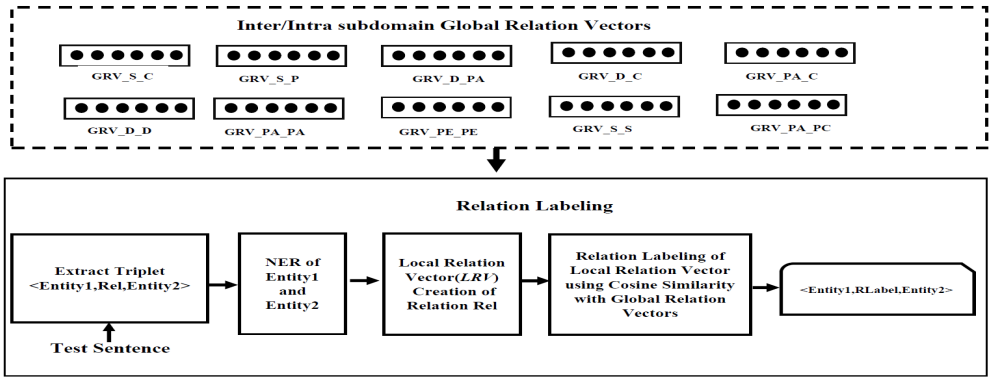


FIGURE 6. Pipeline model demonstrating relation labeling and triplet extraction stage.

it is observed from Table 5, the potential relation patterns among the disease and crop entities are *attacks, is disease in, developing in, is disease affecting, is a common disease of, is a serious disease of*. These relation patterns are vectorized and created a unique Global Relation Vector (*GRV_D_C*) in the inter subdomain <*Disease,Crop*>. Similarly, we created GRVs for the other inter/intra subdomains as mentioned in Table 5. As shown in Figure 6, once the GRVs are defined, we need to identify the Local Relation Vector (LRV)for the relation pattern that is present in the input test sentence. During the testing phase, the test sentence is passed to the OIE system, and a triplet (*Entity*1, *Rel*, *Entity*2) is generated. The arguments Entity1 and Entity2 of the triplet are labeled using our AGRONER module as *l1* and *l2*, respectively as shown in Equation (3). The input relation pattern $input\_rp = \{w_1, w_2 \ldots w_m\}$ is defined as the relation *Rel* present in the triplet (*Entity*1, *Rel*, *Entity*2). The *input_rp* is vectorized and the Local Relation Vector (*LRV*) is generated using Equation (4), where *M* represents the number of words in the input relation pattern, w_i refers to the *i*th word of the relation pattern. By calculating the cosine similarity score between the *LRV* and GRV, we can determine the similarity between them as shown in Equation (5).

Consider the following 5 input sentences:

S1: Brown rot is a common and destructive disease of apricots.
S2: Brown rot is the most serious disease in plums, tart cherries and apricots in Minnesota.
S3: Early blight is a fungal disease caused by Alternaria solani.
S4: Clayey soil and loamy soil are suitable for growing cereals like wheat and gram.
S5: The pesticide yraclostrobin is used to protect Fragaria, Rubus idaeus, Vaccinium corymbosu.

For each sentence, the generated triplets are shown in Table 6. The entities are labeled using the NER module. We can observe from the table that there exists different relation phrases between the same entity pair. The **score** value show the cosine similarity of *Actual Relation with Global Relation Vector GRV*. Based on the cosine similarity score,

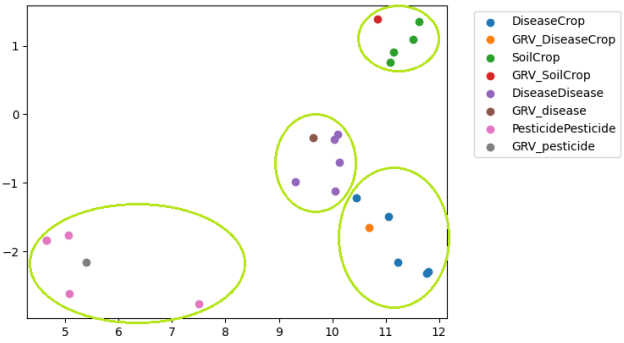


FIGURE 7. Visualization of two inter/intra subdomains.

all the relation patterns are labeled using the label given to the inter subdomains. The final triplets which are used for knowledge graph creation are shown in the last column of Table 6. Similarly, We have created a triplet corpus with 6236 triplets extracted in different inter-intra subdomains. As shown in Equation (6), the label associated with the GRV is assigned as the final relation label if the cosine similarity score exceeds a threshold value $\theta_{\text{subdomain}}$. The threshold $\theta_{\text{subdomain}}$ is determined as the average similarity score of all relation patterns present in an inter subdomain with the GRV. The threshold values ($\theta_{\text{subdomain}}$) assigned for different inter-intra subdomains are as follows: $\theta_{S\_L} = 0.75$, $\theta_{S\_C} = 0.85$, $\theta_{D\_C} = 0.83$ $\theta_{D\_PA} = 0.78$, $\theta_{PA\_C} = 0.79$ $\theta_{PE\_PA} = 0.80$ $\theta_{S\_S} = 0.90$ $\theta_{D\_D} = 0.90$ $\theta_{PE\_PE} = 0.89$ $\theta_{PA\_PA} = 0.87$.

In the Figure 7, visualization of two inter and two subdomains with global vectors are plotted. By referring to Figure 7, it becomes evident that all the relation phrases in the <Disease, Crop> inter subdomain are clustered together, displaying similarity with the *GRV_D_C*.

C. AGRI RELATION EVALUATION AND AGRI TRIPLET VISUALIZATION MODULE

In this section, we present the evaluations of extracted relation patterns and visualization of triplets. To evaluate the effectiveness of our triplet extraction model, we utilized a

TABLE 6. Output of relation labeling and triplet extraction stage.

Sl:No	Triplet	Entity1	Label of Entity1 *l1*	Rel	Entity2	Label of Entity2 *l2*	Score	Predicted Relation (RLabel)	Final Triplet for KG Creation
1	T1	Brown rot	Disease	is common disease of	apricots	Crop	0.8988	Affect	<Brown rot,Affect,Apricots>
	T2	Brown rot	Disease	is destructive disease of	apricots	Crop	0.9029	Affect	<Brown rot,Affect,Apricots>
2	T1	Brown rot	Disease	is most serious disease in	apricots	Crop	0.9282	Affect	<Brown rot,Affect,apricots>
	T2	Brown rot	Disease	is most serious disease in	tart cherries	Crop	0.9282	Affect	<Brown rot,Affect,tart cherries>
	T3	Brown rot	Disease	is most serious disease in plums in	Minnesota	Place	0.9240	OTHER	<Brown rot,OTHER,Place>
3	T1	Early blight	Disease	is fungal disease caused by	Alternaria solani	Pathogen	0.9464	Caused by	<Early blight,Caused by,Alternaria solani>
	T2	Early blight	Disease	is	fungal disease	Disease	0.8407	isa_Disease	<Early blight,isaDisease,Disease>
	T3	Early blight	Disease	be caused by	Alternaria solani	Pathogen	0.8323	Caused by	<Early blight,Caused by,Alternaria solani>
4	T1	Clayey soil	Soil	are suitable for growing cereals like	wheat	Crop	0.9464	suitable for	<Clayey soil,suitable for,wheat>
	T2	Clayey soil	Soil	are suitable for growing cereals like	gram	OTHER	0.9464	OTHER	<Clayey soil,OTHER,Wheat>
	T3	Loamy soil	Soil	are suitable for growing cereals like	wheat	Crop	0.9464	suitable for	<Clayey soil,suitable for,wheat>
	T3	Loamy soil	Soil	are suitable for growing cereals like	gram	OTHER	0.9464	OTHER	<Clayey soil,OTHER,gram>
5	T1	Pesticide yraclostrobin	Pesticide	is used to protect	Fragaria	Pathogen	0.8323	Control	<Pesticide yraclostrobin,Control,Fragaria>
	T2	Pesticide yraclostrobin	Pesticide	is used to protect	Rubus idaeus	Pathogen	0.8323	Control	<Pesticide yraclostrobin,Control, Rubus idaeus >
	T3	Pesticide yraclostrobin	Pesticide	is used to protect	Vaccinium corymbosu	Pathogen	0.8323	Control	<Pesticide yraclostrobin,Control, Vaccinium corymbosu>

TABLE 7. Proposed model's accuracy, precision, recall, and f-measure in labeling inter/intra relations.

Sl:No	Inter/IntraSubdomains	A	P	R	F-Score
1	$\langle Soil, Crop \rangle$	89.89%	94%	92%	93%
2	$\langle Soil, Place \rangle$	92.06%	81%	87%	84%
3	$\langle Pathogen, Crop \rangle$	86.25%	89%	75%	81%
4	$\langle Disease, Pathogen \rangle$	88.89%	93%	89%	91%
5	$\langle Pesticide, Pathogen \rangle$	90.91%	88%	79%	83%
6	$\langle Disease, Crop \rangle$	87.16%	88.0%	92%	90%
7	$\langle Pesticide, Pesticide \rangle$	82.61%	86%	81%	83%
8	$\langle Disease, Disease \rangle$	87.96%	91%	86%	88%
9	$\langle Pathogen, Pathogen \rangle$	89.62%	91%	89%	90%
10	$\langle Soil, Soil \rangle$	91.55%	87%	87%	87%

test corpus containing 3500 agriculture triplets. In Table 7, we present the performance evaluation of the proposed RE model for identifying both inter-subdomain and intra-subdomain relations. The proposed model yielded promising results, achieving an average F Measure of 87%. We performed a detailed inter-intra subdomain analysis, as shown in Table 7. We observed high F-score in the classification of relations belonging to the $\langle Soil, Crop \rangle$ relation, while the lowest F-score was observed for $\langle Pathogen, Crop \rangle$ entities. This performance variation can be primarily attributed to errors originating from the OIE system and unsupervised NER modules.

To visualize the triplets, we utilized the agri triplet corpus comprising 6236 triplets, extracted from the agriculture corpus created using the bootstrapping approach. The data statistics of the triplets are shown in Table 8. There are many different tools and technologies available for creating and visualizing the knowledge graph. We use NetworkX [114], the Python package to create, and analyse the structure of agriculture knowledge graph. We created a MultiDiGraph that holds directed multiedges and self loops between entities. Multiedges refer to the presence of multiple edges between two nodes in a graph, where each edge has the ability to store optional data or attributes. Figures 14 to 15 present the subgraphs generated from agriculture knowledge graph.

D. AGRICULTURE KNOWLEDGE GRAPH (AGRIKG) COMPLETION

In this module the Agriculture Knowledge Graph (AgriKG) is created and information is extracted. The entire process is divided into two stages such as AgriKG Creation and Reasoning (Stage 1) and AgriKG Prediction and Evaluation (Stage 2). The following subsections explain the process in detail.

TABLE 8. Data statistics of the triplets for knowledge graph creation.

Sl:No	Inter/Intra Subdomains	No of Seed Relations	No of New Relations	%Change
1	<Soil,Crop>	25	94	276%
2	<Pathogen,Crop>	150	514	242.6%
3	<Soil,Location>	25	103	312%
4	<Disease, Crop>	150	579	286%
5	<Disease, Pathogen>	1000	1205	20.5%
6	<Pesticide, Pathogen>	50	79	58%
7	<Soil,Soil>	20	43	115%
8	<Pathogen,Pathogen>	1000	1195	19.5%
9	<Disease,Disease>	2100	2300	10%
10	<Pesticide, Pesticide>	100	124	24%

TABLE 9. Presence or absence of properties of different knowledge graph embedding models.

Sl:No	Models	Relation Patterns			
		Sym	Anti-sym	Inv	Comp
1	DistMA	✓	×	×	×
2	ConvE	✓	✓	✓	×
3	HoIE	✓	✓	✓	×
4	CrossE	✓	✓	×	✓
5	TransD	✓	✓	×	×
6	QuatE	✓	✓	✓	×
7	TransR	✓	✓	×	×
8	TransH	✓	✓	×	×
9	BoxE	✓	✓	✓	×
10	TransE	×	✓	✓	✓
11	MuRE	×	✓	✓	✓
12	RotatE	✓	✓	✓	✓
13	TuckER	✓	✓	✓	✓
14	PairRE	✓	✓	✓	✓

1) AGRIKG CREATION AND REASONING STAGE

In this module, as shown in Figure 1, Agriculture Knowledge Grapgh (AgriKG) is created and assessed using both single-hop and multi-hop queries. The data statistics of the triplets, shown in Table 8, are used for AgriKG creation and Reasoning. The Equation (7) is used for relative percentage improvement of newly identified triplets calculation.

$$RelativePercentage = \frac{NewValue - OldValue}{OldValue} * 100 \quad (7)$$

The created knowledge graphs are incomplete in representing a huge amount of real-world facts. To further expand knowledge graphs, Knowledge Graph Embedding (KGE) is utilized. KGE is the process of probabilistically inferring missing connections within knowledge graphs by leveraging the existing facts. One of the key benefits of using KGE in the agriculture domain is that it make connections and discover relationships between different pieces of entities that may not be immediately apparent.

There are different flavours of KGE that have been developed over the course of the past few years. In order to address the knowledge graph incompleteness problem, various Link Prediction (LP) methods are proposed such as ComplEx [90], TorusE [115], ConvE [91], HolE [89], CroosE [116], TransD [117], TransE [83], QuatE [118], RotatE [86], TuckER [94], TransR [119], MuRE [93], TransH [120], BoxE [121], and PairE [95]. All models use score function to compute how distant two nodes relative to its relation type. The score function takes as input the embeddings of the head entity, relation, and tail entity, and produces a real-valued score. This score indicates how well the triple aligns with the underlying semantics of the knowledge graph. For example, TransE, operates on a triple (h, r, t), where h represents the head entity, r represents the relation/predicate, and t represents the tail entity. The objective of TransE is to learn embeddings in such a way that the relation r can be used to translate the head entity h to the tail entity t. It utilizes a scoring function $f(h,t) = ||e_h + r_r - e_t||$ to evaluate the triplets.

TABLE 10. Properties of link prediction models.

Relation	Relation Pattern
Symmetric	$\forall x, y \ \text{r(x,y)} \implies r(y,x)$
Anti-symmetric	$\forall x, y \ \text{r(x,y)} \implies \neg \ r(y,x)$
Inverse	$\forall x, y, r1(x,y) \implies r2(y,x)$
Composition	$\forall x, y, z \ r2(x,y) \wedge r3(y,z) \implies r1(x,z)$

The effectiveness of link prediction methods relies on their capacity to infer various relation patterns, encompassing characteristics such as symmetry, asymmetry, inversion, and composition. These patterns play a crucial role in capturing the complex nature of relationships within a knowledge graph. The formal definitions [86] of these relation patterns are presented in Table 10. However, none of the current approaches can cover them well. In Table 9, properties of Link Prediction Models are presented. Some of the recent KGE models, along with score functions, are described in Table 16.

The proposed Agriculture KG is evaluated using rank-based metrics viz., Hits@N, Mean Reciprocal Rank, and Mean Rank, which are computed using the equations (8), (9), and (10). In all these measures, positive triplets are scored against negative ones to determine whether the model is able to predict plausible facts. In Equations, Q refers the batch of ground-truth triplets and *rank()* is the position of the ground-truth triple in the sorted list of triplets. Triplets can have the form (h, r, t). The hits@N [122] describes the fraction of true entities that appear in the first N entities of the sorted rank list. It lies between (0,1)where closer to 1 is better.

$$Hits@N = \sum_{i=1}^{|Q|} 1 \text{ if rank}(h,r,t) \leq N \quad (8)$$

The Mean Reciprocal Rank (MRR) [123] is a relative score that calculates the average or mean of the inverse of the ranks at which the first relevant document was retrieved for a set of queries. MRR is bound on (0,1) where closer to 1 is better.

$$MeanReciprocalRank = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank(h,r,t)i} \quad (9)$$

The mean rank (MR) [123]computes the arithmetic mean over all individual ranks. The MR lies on the interval

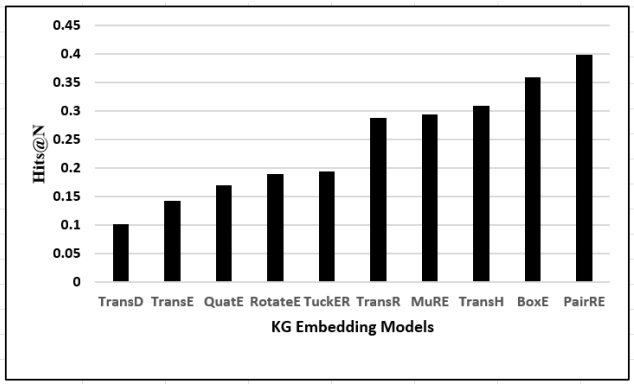


FIGURE 8. Hits@10 of different knowledge graph embedding models.

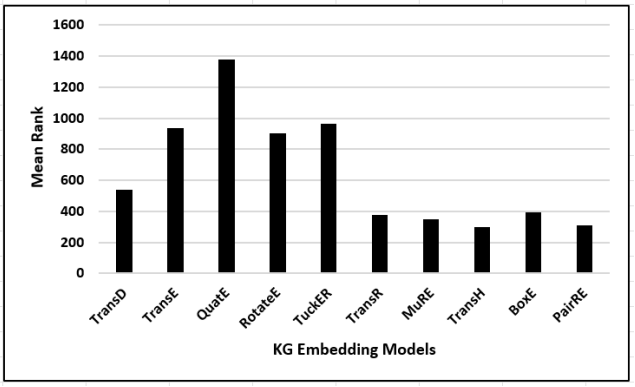


FIGURE 9. Mean rank of different knowledge graph embedding models.

MR $\in (1,\infty)$ where lower is better.

$$MeanRank = \frac{1}{|Q|}\sum_{i=1}^{|Q|} rank_(h, r, t)i \tag{10}$$

We perform our experiments to latest Knowledge Graph Embedding (KGE) models viz., TransD TransE, QuatE, RotatE, TuckER, TransR, MuRE, TransH, BoxE, and PairRE. We used Pykeen (Python KnowlEdge EmbeddiNgs)[5] for the implemenation of KGE models. These models are evaluated using Equations 8, 9, and 10. Based on our experiments we choose top 5 KGE Models, their associated Hits@10 are shown in Figure 8. It can be noticed that the PairRE model presents the highest hits@10. In Figure 9, Mean Rank of KG Embedding models are presented. It is found that the least value is for PairRE. In Figure 10, the Mean Reciprocal Rank of KGE models are presented. We noticed that PairRE and TransH are giving highest score. Further analysis was carried out to assess the efficiency of this model for knowledge graph reasoning. We tested the top five models with one-hop queries and multi-hop queries. In one-hop queries, the query can have the form (?,r,t),(h,r,?), (h,r,?) where **(h, r, t)** represents the head entity, relation and tail entity respectively. The missing entity or relation is represented by *"?"*. The head entity prediction for the queries **Query('?','causedby','Cercospora mamaonis')**, and **Query('?',suitable for','alluvial soil')** by the KGE models are presented in Table 11. The relation prediction for the queries **Query('Fruit spot','?','Cercospora mamaonis')** and **Query('Bajra', '?', 'Alluvial Soil'** by the KGE models are presented in the Table 12. In each table, the correct predictions are emphasized by being highlighted in bold. Given the **Query('Fruit spot ','causedby',?)**, the model predictions for tail entities are shown in Table 13. We noticed that majority of the models are good at predicting one hop queries which identifies head, tail entities and relations.

[5]https://pykeen.readthedocs.io/en/stable/

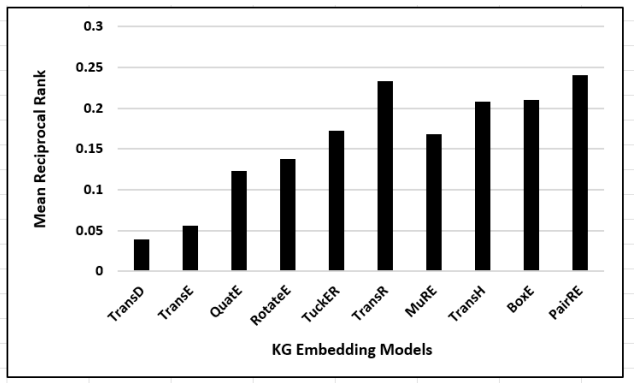


FIGURE 10. Mean reciprocal rank of different knowledge graph embedding models.

2) AGRIKG PREDICTION AND EVALUATION

After employing rank-based metrics like Hits@N, MRR, and MR, we observed that the PairRE model outperformed other Knowledge Graph Embedding (KGE) models. Consequently, we proceeded to evaluate the performance of the PairRE model on both one-hop and multi-hop queries for relation prediction. Twenty random queries were selected from both inter and intra subdomains. Figure 11 displays the comprehensive performance results of the PairRE model when handling one-hop queries. The PairRE model achieves a an average F-score of 92% in the identification of ten inter-intra relations. For a more in-depth understanding, we conducted an individual analysis of relations, as shown in Figure 12.

We also tested PairRE model with multihop queries. For instance the triplet('Apple mosaic','caused by','Ilarvirus') and triplet('Apple mosaic', 'affect', 'Apple') are present in the training set and the query is ('apple',?,'Ilarvirus'). After multiple experiments with the top five KGE models, we have identified that PairRE demonstrates good performance when handling mutltihop queries. The top 5 predictions of the KGE models are presented in Table 14. In spite of the TuckER model's precise predictions, we have observed that the PairRE model correctly predicts the output for 20 randomly selected queries. In the table 14, the accurate predictions

TABLE 11. Top five head entity prediction: one hop query.

Sl:No	Models	Top 5 Predicted Head Entities
		Query('?','causedby','Cercospora mamaonis')
1	TransE	**Fruit spot**, Soybean dwarf, Cercospora mamaonis, Pythium stalk rot, Cryptomeria
2	MuRE	**Fruit spot**, Rice White mold, Pythium stalk rot, Brown spot
3	RotatE	**Fruit spot**, disease Potato wart, Rhizoctonia ear rot, Peters scorch
4	TuckER	**Fruit spot**, Oat blue dwarf Virginiacreeper, Pythium stalk rot, Septoria speckled leaf blotch
5	PairRE	**Fruit spot**, Coffee rust, Sappy bark Papery bark, Bacterial leaf spot,Angular leaf spot
		Query('?',suitable for','alluvial soil')
1	TransE	Alluvial soil, **barley** **Jowar**, Synchytrium root gall,**black gram**
2	MuRE	**Jute**, Millets, **Corn**, Coconut,**Groundnut**
3	RotatE	**Bajra,black gram, soyabean, jowar, barley**
4	TuckER	**green gram, bajra soyabean, barley, black gram**
5	PairRE	**Bajra, Wheat, green gram, barley jowar**

TABLE 12. Top five relation prediction:one hop query.

Sl:No	Models	Top 5 Predicted Relations
		Query('Fruit spot', '?', 'Cercospora mamaonis')
1	TransE	**causedby**, isachemical, infect, affect, isaplace
2	MuRE	**causedby**, isadisease, isacrop, suitablefor, infect
3	RotatE	**causedby**, infect, control, isaplace, isapathogen
4	TuckER	**causedby**, isapathogen, isacrop, control, infect
5	PairRE	**causedby**, suitablefor, infect, isacrop, isapathogen
		Query('Bajra', '?', 'Alluvial Soil'
1	TransE	**Suitable for**, found,isa_pathogen, infect, affect
2	MuRE	**suitable for**, isa_disease, isa_soil, infect, found
3	RotatE	**suitable for**,caused by,isa_disease, isa_chemical,isa_soil
4	TuckER	**suitable for**,found, isa_pathogen,isa_chemical,control
5	PairRE	**Suitable for**, control, isa_soil, infect, isa_chemical

are marked with bold text. The PairRE model's overall performance in handling multi-hop queries is reported in Figure 13. We have observed that the precision in recognizing

TABLE 13. Top five tail entity prediction: one hop query.

Sl:No	Models	Top 5 Predicted Tail Entities
		Query('Fruit spot', 'causedby', '?')
1	TransE	**Cercospora mamaonis**, Cyazofamid, Fruit spot, Ampestris, Aphid blight
2	MuRE	Septoria pisi, Corynespora cassicola, **Cercospora mamaonis**, Cochliobolus Thielaviopsis basicola
3	RotatE	**Cercospora mamaonis**, disease, Elm stripe, Phanerogamic root Streptomyces
4	TuckER	**Cercospora mamaonis** Citrus viroid VI, Botryosphaeria stem canker, Pepper veinal mottle
5	PairRE	**Cercospora mamaonis**, Exocortis, red clover, Erwinia cypripedii, Dwarf virus
		Query('Bajra','suitable for','?')
1	TransE	**Alluvial soil**, barley Jowar, Synchytrium root gall,black gram
2	MuRE	**Alluvial Soil**, Black Cotton Soil Hill Soil, groundnuts, Laterite soil
3	RotatE	**Alluvial soil**, Cactus cyst, Rhizoctonia Melanosis ,Pleospora leaf spot
4	TuckER	**Alluvial soil**, Vascular wilt , clay loamy soils Iturin A , Citrus bark cracking viroida
5	PairRE	**Alluvial soil**, Clayey soil, Peaty soil, black cotton soil

TABLE 14. Models and top five multi-hop relation prediction.

Sl:No	Models	Top 5 Predicted Relations
		Query('Apple',?,'Ilarvirus')
1	TransE	affect, control, isasoil, causedby, found
2	MuRE	affect, isasoil, isadisease, suitablefor, caused by
3	RotaE	isapathogen, suiablefor, affect, isasoil, isadisease
4	TuckER	caused by, found, isapathogen, isasoil, suitable for
5	PairRE	**infect**, isasoil,control, suitablefor, isachemical
		Query('Walnut',?,'Xanthamona campestris')
1	TransE	affect,**infect**, isa_pesticide, found, causedby
2	MuRE	isa_disease, isa_soil, infect, affect,isa_pesticide
3	RotaE	isa_disease ,**infect** , suitable for, caused by, affect
4	TuckER	**infect**,isa_pathogen, caused by, affect,isa_disease
5	PairRE	control **infect**, isa_soil,suitable for, found

multi-hop relations is lower compared to other one-hop queries. Consequently, this has led to a decline in the F score.

IV. DISCUSSION AND LIMITATION

We conducted a thorough analysis of the errors generated by the proposed System. The limitations of the proposed model used for knowledge graph creation and reasoning are discussed in this section. The triplet extraction process used by the proposed model relies on the bootstrapping approach. For our experiments we utilize an agriculture corpus of 30k sentences. A large corpus is essential for training and developing effective machine learning models. Certainly, improvements will be made when a substantial corpus becomes available. The model also utilizes Dependency Parsing, OIE system and unsupervised NER models.

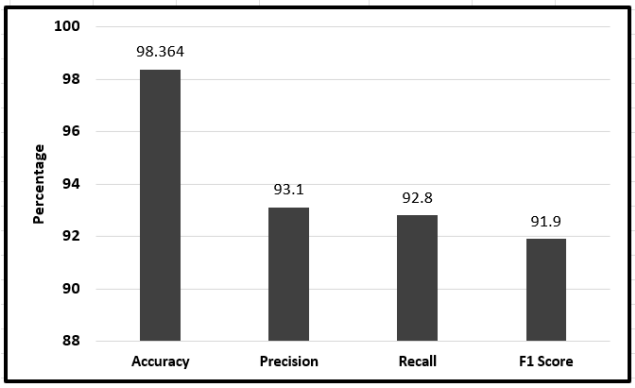


FIGURE 11. Average of Accuracy, Precision, Recall, and F-Measure of the PairRE Model for evaluating one-hop queries.

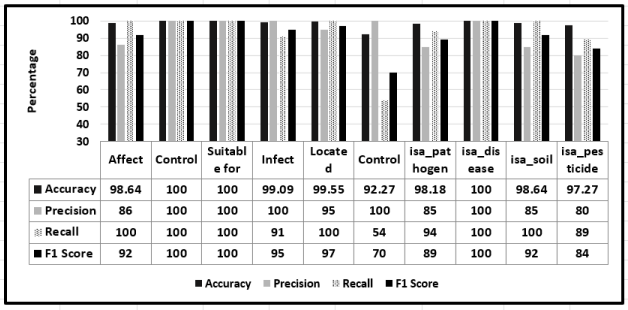


FIGURE 12. Accuracy, Precision, Recall, and F-Measure of the PairRE Model in Recognizing Ten inter/intra Relations.

Therefore, we will discuss the limitations associated with all of these aspects.

1) ERROR GENERATED BY THE DEPENDENCY PARSER

Dependency parsing, like any natural language processing task, can produce errors due to the complexity and ambiguity of natural language. Some major errors generated by dependency parsing include:

- Parsing long-distance dependencies, where a word's relationship to another word is far away in the sentence, can be challenging and prone to errors.
- If the input sentence contains grammatical errors or is poorly structured, dependency parsers may have difficulty producing accurate parses.
- Dependency parsers rely on pre-trained models with limited vocabularies. When encountering out-of-vocabulary words, errors can occur.
- In domain-specific or technical text, parsers may not be as accurate because they are trained on more general language data.

2) ERROR GENERATED BY THE UNSUPERVISED NER MODEL

We have identified the following factors that contribute to the performance decrease in the NER model:

- The absence of entries in the LDA dictionaries utilized for storing words and weights, absence of certain tokens in the tokenizer, missing vectors in word embedding dictionaries, and substantial overlap of concepts/tokens across different subdomains like disease and pathogen.
- Model focuses on identifying six main agricultural entities, namely disease pathogen, pesticide, crops, soil, and place, in the agriculture domain. It classifies other entities that are relevant to agriculture, such as agricultural inputs, farm machinery, livestock animals, agricultural organizations, agricultural technologies, climate zones, and agricultural practices, as 'OTHER'.

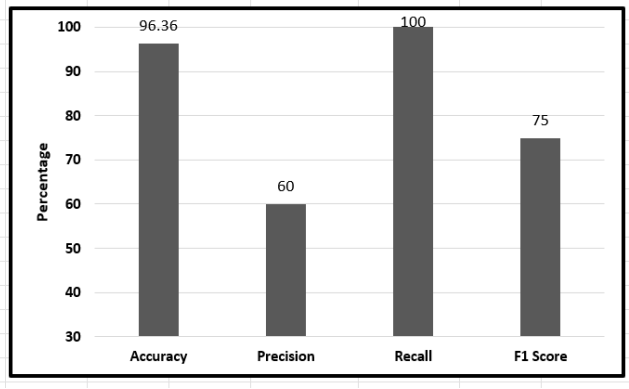


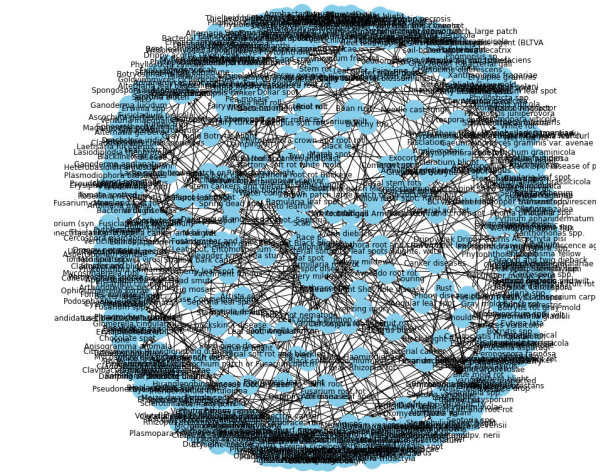
FIGURE 13. Average of Accuracy, Precision, Recall, and F-Measure of the PairRE Model for evaluating multi-hop (infect) queries.

TABLE 15. Details of LDA parameters.

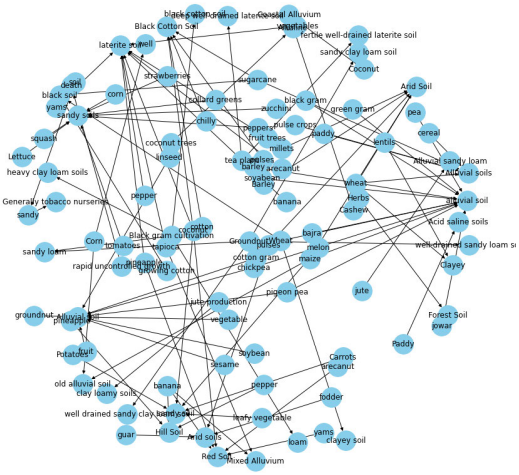
Domain	Document-topic density	Topic-word density	T	K	Number of word-weight distributions in LDA dictionary
Pesticide	[0.062 0.068 0.042]	0.33	3	30	400
Soil	[0.038 0.096 0.033 0.050 0.047 0.055]	0.17	6	30	200
Pathogen	[0.134 0.072 0.119 0.064 0.080 0.199]	0.17	6	30	500
Disease	[0.032 0.052 0.044 0.089 0.052 0.046 0.042]	0.14	7	30	675

3) ERROR GENERATED BY THE BOOTSTRAPPING BASED RELATION EXTRACTION MODULE

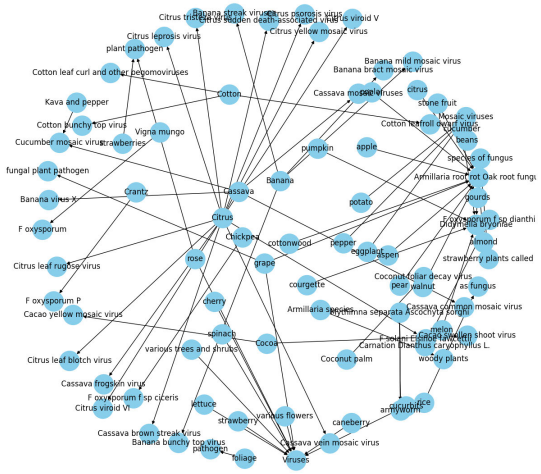
We observed that the bootstrapping approach can produce incorrect relation phrases when the gap between entities within a sentence becomes extensive. In such cases, the approach may struggle to establish a direct connection or identify the appropriate relation pattern between the entities. Given the seed entity('Red Soil, Soil), the model generates non coherent relation **appears because presence the** from the sentence **Red soil appears red because of the presence of iron in its oxides form in the soil**. These relation phrases may not always capture the desired relations accurately, leading to noisy training and potentially lower performance.



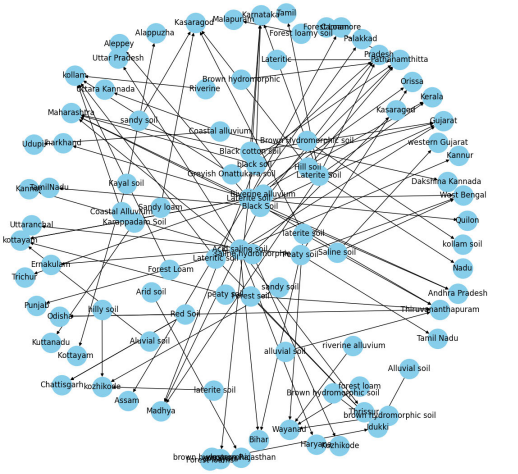
[A]



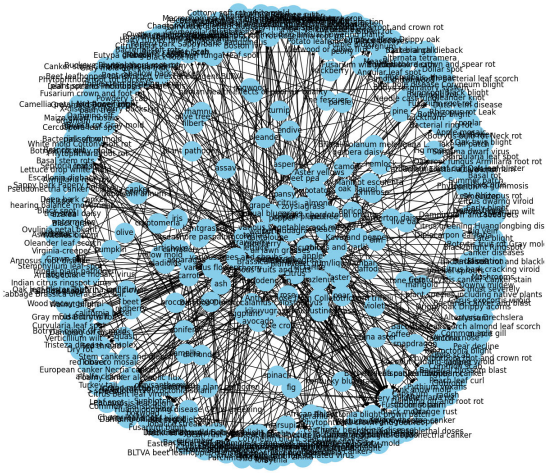
[B]



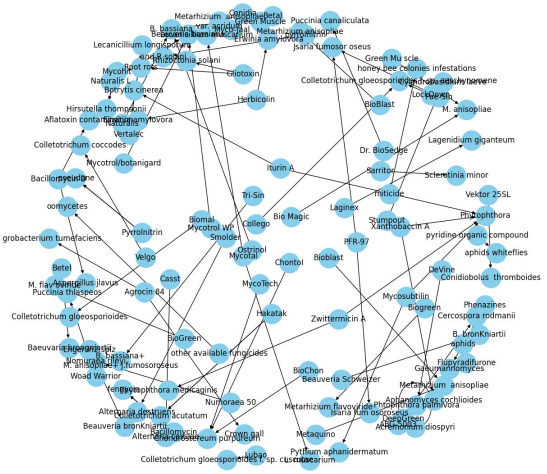
[]]



[D]



[E]



[F]

FIGURE 14. [A]:A subgraph showing *caused by* relation between Disease and Pathogen.[B]:A subgraph showing *Suitable for* relation between Crop and Soils. [C]: A subgraph showing *infect* relation between Pathogen and Crop. [D]:A subgraph showing *found* relation between Soil and Location [E]:A subgraph showing *Affect* relation between Disease and Crop [F]:A subgraph showing *Control* relation between Pesticides and Pathogens.(Inter Subdomain Subgraphs)(Intra Subdomain Subgraphs).

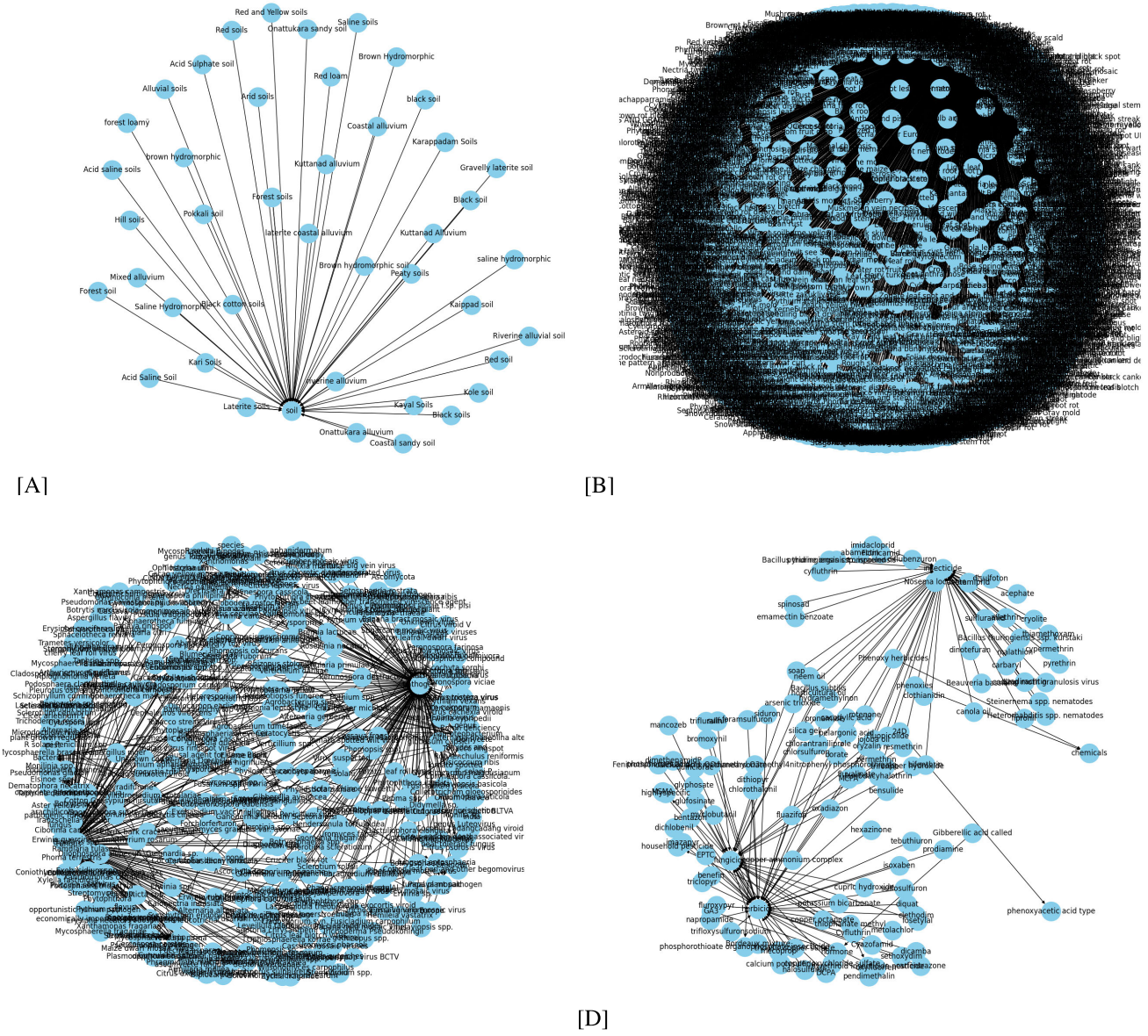


FIGURE 15. **[A]:A subgraph showing *isa* relation in Soil domain.[B]:A subgraph showing *isa* relation in Disease domain. [C]: A subgraph showing *isa* relation in Pathogen domain [D]: A subgraph showing *isa* relation in Pesticide domain. (Intra Subdomain Subgraphs).**

4) ERROR GENERATED BY THE OPEN INFORMATION EXTRACTION SYSTEM

We utilize the OIE system to enhance the relation extraction model. When handling complex and compound sentences, the OIE system tends to produce inaccurate entities. The common errors are listed below:

- Consider an example sentence: *Potato late blight caused by Phytophthora infestans is the most devastating disease of potatoes*. The triplet *"Potato late blight caused by Phytophthora infestans," "is most devastating disease of," "potatoes"* is identified by the OIE system. Potential issues with entity generation are highlighted by the fact that the first argument of this triplet **"Potato late blight caused by Phytophthora infestans"** includes both disease and pathogen entities, which may cause misclassifications by the NER Module.
- One limitation of the triplet extraction module is that it treats related entities as separate entities, leading to the generation of different nodes in the knowledge graph. For example, entities like ''brown rot'' and ''brownrot infection'' are considered two distinct entities, resulting in the creation of two separate nodes in the knowledge graph.

5) MISSING INTER SUBDOMAIN RELATIONS

The proposed AgriKG model missed the following inter subdomain relations in the agriculture domain, such as

TABLE 16. Score function of knowledge graph embedding models.

Sl:No	Models	Score Function	Explanation
1	TransD	$f(h,r,t) = \|\mathbf{M_r} \cdot \mathbf{h_e} + \mathbf{r} - \mathbf{M'_r} \cdot \mathbf{t_e}\|_2$	$\mathbf{M_r}, \mathbf{M'_r}$ are the projection matrices associated with the relation r. $\mathbf{h_e}, \mathbf{t_e}$ are the entity embeddings for h and t. The projected embeddings of the head entity, the relation, and the tail entity are used to compute score function.
2	TransE	$f(h,r,t) = \|h + r - t\|$	Given triplet (head entity, relation, tail entity), the embedding of the head entity (h) is translated by the embedding of the relation (r). The result vector represents the expected position of the tail entity.
3	QuatE	$f(h,r,t) = \text{Re}\left(\frac{h \otimes r}{\|h \otimes r\|} \otimes \bar{t}\right)$	The score function in QuatE calculates the quaternion product of h and r and normalizes it by dividing it by its quaternion norm. Then, the resulting quaternion is multiplied by the quaternion conjugate of t and its real part is taken as the final score.
4	RotatE	$f(h,r,t) = -\text{Re}(\exp(i\theta) \cdot (h \odot r \odot \bar{t}))$	The score function in RotateE calculates the element-wise product of the head entity h, relation r, and the complex conjugate of the tail entity t. Then, it applies the phase angle associated with the relation to the resulting complex number using the exponential function. Finally, the score is the taken as the real part of the complex number.
5	TuckER	$f(h,r,t) = \text{sigmoid}(W \times_1 h \times_2 r \times_3 t)$	The Tensor factorization method is utilized in this model. The score function calculates the tensor contraction of the embeddings of the head entity, the relation, and the tail entity, along with the weight tensor W .
6	TransR	$f(h,r,t) = \|\mathbf{M_r} \cdot \mathbf{h_e} + \mathbf{r} - \mathbf{t_e}\|_2$	In the equation, the $\mathbf{h_e}$ and $\mathbf{t_e}$ are the projected embeddings of h and t in the entity space. $\mathbf{M_r}$ is the projection matrix that maps the entity embeddings from the entity space to the relation space. The score function computes the Euclidean distance among the projected embeddings of the head entity, the relation, and the tail entity.
7	MuRP	$\phi_{\text{MuRP}}(hs, r, ho) = -d(h_s^r, h_t^r)^2 + b_s + b_o$	Multi-Relational Poincaré model utilizes Möbius matrix-vector multiplication and Möbius addition to learn relation-specific parameters. In the score functoion h_s^r, h_t^r are the hyperbolic embeddings of the head and tail entities.
8	TransH	f(h, r, t) = $-\|e'_{h,r} + d_r - e'_{t,r}\|_p^2$	In the score function, $e'_{h,r}$, $e'_{t,r}$ are the projected embeddings of h and t onto the hyperplane defined by the relation r.
9	BoxE	$f(h,r,t) = \sum_{i=1}^{n} \|\text{dist}(e_i^{r(e_1,\ldots,e_n)}, r^i\|_x$	In this model entities are embedded as points, and relations are considered as a set of boxes. The embedding of the entity e_i is represented as $e_i^{r(e_1,\ldots,e_n)}$ and r^i is the hyper-rectangle associated with r. A distance function is used for evaluating entity positions relative to the box positions.
10	PairE	f(h, r, t) = $\|h \circ r^H - t \circ r^T\|$	In the score function, the head entity h is projected to Euclidean space using vector r^H, while the tail entity t is projected using vector r^T. The projection operation is performed through the Hadamard product, which involves element-wise multiplication of these two vectors.

<Disease, Place>, <Disease, Pesticide>, and <Crop, Pesticides>. Consequently, the existing model has been evaluated using a single type of multihop relational query, specifically involving the concept of *infect*. In order to uncover additional similar relationships, it is required to introduce a greater number of inter-subdomain triplets.

6) MISLABELING OF RELATIONS BY THE PROPOSED MODEL

The proposed model aims to identify six inter and four intra subdomain relations in the agriculture domain. However, several other possible relations in domain, such as (Crop, Fertilizers), (Crop, Climate), (Soil, Fertilizers), (Crop, Pesticides), (Disease, Pesticides), and (Disease, Place). These relations are currently labeled as 'OTHER' by the model. An example of the (Disease, Place) relation mislabeled as 'OTHER' can be found in Table 6.

V. CONCLUSION AND FUTURE WORKS

The approach proposed offers a semi-supervised syntax-semantics based approach for agriculture relation extraction. These extracted triplets are subsequently utilized in creating and reasoning about an AgriKG. This approach addresses

the challenges of creating domain-specific knowledge graphs, such as domain-specific vocabulary, data integration challenges, dynamic data, and the need for domain expertise. The proposed AgriKG emphasis on 10 inter-intra relationships, viz,. *<Soil,Location>, <Soil,Crop>, <Disease, Pathogen>,<Pathogen,Crop>,<Pesticide,Pathogen>, <Disease,Crop>, <Disease,Disease>, <Pathogen,Pathogen>, <Soil,Soil>, and <Pesticide,Pesticide>*. To perform entity identification we use OIE system, while extended BERT with LDA was used for NER. By employing dependency parsing techniques and advanced NER models based on BERT and LDA, this approach improves the accuracy of triplet extraction and entity recognition in the agricultural domain. Our approach introduces semi-supervised bootstrapping that leverages the dependency parse tree to identify relation patterns and entities for triplet generation. BERT based relation classifier is also employed in this approach to classify new relation patterns. The proposed approach for triplet extraction achieved a macro F score of 87%. The creation of a benchmark dataset and the achievement of a high F-score for relation extraction demonstrate the effectiveness of this approach in capturing the relationships between entities. The knowledge graph that has been created can be used as a valuable resource for addressing various inquiries raised by farmers related to crop diseases, effective soil management techniques, guidelines for pesticide usage, and the latest farming practices.

The proposed approach for knowledge graph reasoning focuses on answer reasoning with binary facts. In our upcoming projects, we plan to focus on exploring n-ary facts that involve more than two entities. We also plan to focus on more inter subdomain relations in the agriculture domain, such as <Disease, Place>, <Disease, Pesticide>, and <Crop, Pesticides> and more multi-hop query prediction. In our future works, we intend to enhance the model through further fine-tuning of large language models. We plan to use Large Language Models (LLMs) in AGRONER model to generate additional content, explanations, or context for topics identified by LDA. In place of BERT plan to experiment with several transformer-based alternatives such as RoBERTa, ALBERTA, XLNet, etc.

APPENDIX DETAILS OF KEY PARAMETERS

In our experiments we used the following hyper parameters:

- α: This hyperparameter controls the document-topic density.
- η: This hyperparameter controls the topic-word density.(eta in Gensim[6]
- Number of iterations.

The other parameters are:

- T: Number of topics required.
- K: Numbers of words required in a topic.
- Number of word-weight distributions required to create the dictionaries for Global/Local Vector creation.

[6]https://pypi.org/project/gensim/ Library)

Details of LDA parameters are presented in the following Table 15.

REFERENCES

[1] S. S. Kondekar, "An analytical note on the status of agriculture sector in Indian economy," *PARIPEX Indian J. Res.*, vol. 12, pp. 25–27, Jan. 2023.

[2] K. K. Kumar, K. R. Kumar, R. G. Ashrit, N. R. Deshpande, and J. W. Hansen, "Agriculture role on Indian economy," *Bus. Econ. J.*, vol. 6, no. 4, 2015.

[3] R. M. Bhise, "Agriculture sector of Indian economy," *Golden Res. Thought*, vol. 5, no. 9, 2016.

[4] K. K. Kumar, K. R. Kumar, R. G. Ashrit, N. R. Deshpande, and J. W. Hansen, "Climate impacts on Indian agriculture," *Int. J. Climatol.*, vol. 24, no. 11, pp. 1375–1393, 2004.

[5] K. S. K. Kumar and J. Parikh, "Indian agriculture and climate sensitivity," *Global Environ. Change*, vol. 11, no. 2, pp. 147–154, Jul. 2001.

[6] J. L. Martinez-Rodriguez, A. Hogan, and I. Lopez-Arevalo, "Information extraction meets the semantic web: A survey," *Semantic Web*, vol. 11, no. 2, pp. 255–335, 2020.

[7] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticæ Investigationes. Int. J. Linguistics Lang. Resour.*, vol. 30, no. 1, pp. 3–26, Aug. 2007.

[8] N. Bach and S. Badaskar, "A review of relation extraction," *Literature Rev. Lang. Statist. II*, May 2011.

[9] X. Hao, Z. Ji, X. Li, L. Yin, L. Liu, M. Sun, Q. Liu, and R. Yang, "Construction and application of a knowledge graph," *Remote Sens.*, vol. 13, no. 13, p. 2511, 2021.

[10] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[11] D. N. Nicholson and C. S. Greene, "Constructing knowledge graphs and their biomedical applications," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1414–1428, Jun. 2020, doi: 10.1016/j.csbj.2020.05.017.

[12] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, Y. Xiao, and N. J. Yuan, "Multi-modal knowledge graph construction and application: A survey," *IEEE Trans. Knowl. Data Eng.*, pp. 1–20, 2022, doi: 10.1109/TKDE.2022.3224228.

[13] R. Navigli, "BabelNet 3.0: A core for linguistic linked data and NLP," in *Proc. 12th Summer School Linguistic Linked Open Data, EUROLAN Workshop Social Media Web Linked Data*, in Communications in Computer and Information Science, vol. 588, 2016.

[14] T. Rebele, F. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum, "YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames," in *Proc. Int. Semantic Web Conf.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 9982, 2016, pp. 177–185.

[15] J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum, "YAGO2: Exploring and querying world knowledge in time, space, context, and many languages," in *Proc. 20th Int. Conf. Companion World Wide Web*, Mar. 2011, pp. 229–232.

[16] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2008, pp. 1247–1250.

[17] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[18] B. Abu-Salih, "Domain-specific knowledge graphs: A survey," 2021, *arXiv:2011.00235*.

[19] L. Murali, G. Gopakumar, D. M. Viswanathan, and P. Nedungadi, "Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study," *J. Biomed. Informat.*, vol. 143, Jul. 2023, Art. no. 104403.

[20] K. Zhang and J. Liu, "Review on the application of knowledge graph in cyber security assessment," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 768, no. 5, Mar. 2020, Art. no. 052103.

[21] M. Hou, R. Wei, L. Lu, X. Lan, and H. Cai, "Research review of knowledge graph and its application in medical domain," *Jisuanji Yanjiu yu Fazhan/Computer Res. Develop.*, vol. 55, no. 12, pp. 2587–2599, 2018.

[22] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000.

[23] D. Zhang, Z. Liu, W. Jia, H. Liu, and J. Tan, "A review on knowledge graph and its application prospects to intelligent manufacturing," *J. Mech. Eng.*, vo. 57, no. 5, pp. 90–113, 2021.

[24] X. Ma, "Knowledge graph construction and application in geosciences: A review," *Comput. Geosci.*, vol. 161, Apr. 2022, Art. no. 105082.

[25] H. Qin and Y. Yao, "Agriculture knowledge graph construction and application," *J. Phys., Conf. Ser.*, vol. 1756, no. 1, Feb. 2021, Art. no. 012010.

[26] Q. Guo, S. Cao, and Z. Yi, "A medical question answering system using large language models and knowledge graphs," *Int. J. Intell. Syst.*, vol. 37, no. 11, pp. 8548–8564, Nov. 2022.

[27] V. G., V. Kanjirangat, and D. Gupta, "AGRONER: An unsupervised agriculture named entity recognition using weighted distributional semantic model," *Expert Syst. Appl.*, vol. 229, Nov. 2023, Art. no. 120440. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417423009429

[28] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[29] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, and Z. Duan, "Knowledge graph completion: A review," *IEEE Access*, vol. 8, pp. 192435–192456, 2020.

[30] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in *Proc. 16th Conf. Comput. Linguistics*, 1996.

[31] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," Tech. Rep., 2015.

[32] S. Bird, "NLTK: The natural language toolkit," in *Proc. 21st Int. Conf. Comput. Linguistics, 44th Annu. Meeting Assoc. Comput. Linguistics, Interact. Presentation Sessions (COLING/ACL)*, 2006.

[33] *Apache OpenNLP*, Apache, 2016. [Online]. Available: https://opennlp.apache.org/

[34] S. Vergara, M. El-Khouly, M. El Tantawi, S. Marla, and S. Lak, "Building cognitive applications with IBM Watson services: Volume 7 natural language understanding," Tech. Rep., 2017, vol. 2.

[35] F. Dernoncourt, J. Y. Lee, and P. Szolovits, "NeuroNER: An easy-to-use program for named-entity recognition based on neural networks," in *Proc. EMNLP Syst. Demonstrations*, Copenhagen, Denmark, Sep. 2017, pp. 97–102.

[36] P. Goyal, S. Pandey, and K. Jain, "SpaCy," in *Deep Learning for Natural Language Processing: Creating Neural Networks With Python*, 2018.

[37] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "AllenNLP: A deep semantic natural language processing platform," 2019, *arXiv:1803.07640*.

[38] A. Mishra, "Amazon comprehend," in *Machine Learning in the AWS Cloud*, 2019.

[39] D. Farmakiotou, V. Karkaletsis, J. Koutsias, G. Sigletos, C. D. Spyropoulos, and P. Stamatopoulos, "Rule-based named entity recognition for Greek financial texts," in *Proc. Workshop Comput. Lexicography Multimedia Dictionaries (COMLEX)*, 2000, pp. 75–78.

[40] G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, "Using machine learning to maintain rule-based named-entity recognition and classification systems," in *Proc. 39th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 426–433.

[41] N. Abinaya, N. John, H. B. B. Ganesh, M. A. Kumar, and K. P. Soman, "AMRITA-CEN@FIRE-2014: Named entity recognition for Indian languages using rich features," in *Proc. ACM Int. Conf. Ser.*, vols. 5–7, 2014, pp. 103–111.

[42] G. Szarvas, R. Farkas, and A. Kocsor, "A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms," in *Proc. Int. Conf. Discovery Sci.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 4265, 2006, pp. 267–278.

[43] V. Krishnan and C. D. Manning, "An effective two-stage model for exploiting non-local dependencies in named entity recognition," in *Proc. 21st Int. Conf. Comput. Linguistics, 44th Annu. Meeting Assoc. Comput. Linguistics (COLING/ACL)*, vol. 1, 2006, pp. 1121–1128.

[44] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[45] W. Li, Y. Guo, B. Wang, and B. Yang, "Learning spatiotemporal embedding with gated convolutional recurrent networks for translation initiation site prediction," *Pattern Recognit.*, vol. 136, Apr. 2023, Art. no. 109234.

[46] Y. Guo, D. Zhou, X. Ruan, and J. Cao, "Variational gated autoencoder-based feature extraction model for inferring disease-miRNA associations based on multiview features," *Neural Netw.*, vol. 165, pp. 491–505, Aug. 2023.

[47] S. Yang, D. Zhou, J. Cao, and Y. Guo, "Rethinking low-light enhancement via transformer-GAN," *IEEE Signal Process. Lett.*, vol. 29, pp. 1082–1086, 2022.

[48] A. Gopalakrishnan, K. P. Soman, and B. Premjith, "A deep learning-based named entity recognition in biomedical domain," in *Proc. Int. Conf. ICERECT*. Singapore: Springer, 2018, pp. 517–526.

[49] K. Jayaram and K. Sangeeta, "A review: Information extraction techniques from research papers," in *Proc. Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Feb. 2017, pp. 56–59.

[50] R. R. K. Menon, D. Joseph, and M. R. Kaimal, "Semantics-based topic inter-relationship extraction," *J. Intell. Fuzzy Syst.*, vol. 32, no. 4, pp. 2941–2951, Mar. 2017.

[51] S. Mellace, K. Vani, and A. Antonucci, "Relation clustering in narrative knowledge graphs," Nov. 2020, *arXiv:2011.13647*.

[52] V. Hariharan, M. Anand Kumar, and K. P. Soman, "Relation extraction using convolutional neural networks," in *Proc. Int. Conf. ISMAC Comput. Vis. Bio-Eng.*, in Lecture Notes in Computational Vision and Biomechanics, vol. 30, 2019, pp. 937–944.

[53] A. Ben Abacha and P. Zweigenbaum, "A hybrid approach for the extraction of semantic relations from MEDLINE abstracts," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 6609, 2011, pp. 139–150.

[54] A. Sharma, R. Swaminathan, and H. Yang, "A verb-centric approach for relationship extraction in biomedical text," in *Proc. IEEE 4th Int. Conf. Semantic Comput. (ICSC)*, Sep. 2010, pp. 377–385.

[55] V. Kanjirangat and F. Rinaldi, "Enhancing biomedical relation extraction with transformer models using shortest dependency path features and triplet information," *J. Biomed. Informat.*, vol. 122, Oct. 2021, Art. no. 103893.

[56] K. Fundel, R. Küffner, and R. Zimmer, "RelEx—Relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, Feb. 2007.

[57] F. Li, M. Zhang, G. Fu, and D. Ji, "A neural joint model for entity and relation extraction from biomedical text," *BMC Bioinf.*, vol. 18, no. 1, pp. 1–11, Dec. 2017.

[58] Y. Zhang, H. Lin, Z. Yang, J. Wang, S. Zhang, Y. Sun, and L. Yang, "A hybrid model based on neural networks for biomedical relation extraction," *J. Biomed. Informat.*, vol. 81, pp. 83–92, May 2018.

[59] A. Sun, R. Grishman, and S. Sekine, "Semi-supervised relation extraction with large-scale word clustering," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol. (ACL-HLT)*, vol. 1, 2011.

[60] L. Tai, F. Guo, and S. Qin, "Semi-supervised entity relation extraction based on trigger word," in *Proc. 3rd IEEE Int. Conf. Comput. Commun. (ICCC)*, Dec. 2017, pp. 497–501.

[61] S. Brin, "Extracting patterns and relations from the world wide web," in *Proc. Int. Workshop World Wide Web Databases*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 1590, 1999, pp. 172–183.

[62] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proc. 5th ACM Conf. Digit. Libraries*, Jun. 2000, pp. 85–94.

[63] D. S. Batista, B. Martins, and M. J. Silva, "Semi-supervised bootstrapping of relationship extractors with distributional semantics," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 499–504.

[64] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2011, pp. 1535–1545.

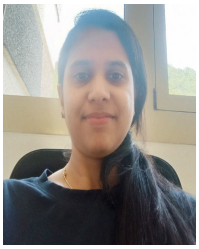
[65] T. Zhu, H. Wang, J. Yu, X. Zhou, W. Chen, W. Zhang, and M. Zhang, "Towards accurate and consistent evaluation: A dataset for distantly-supervised relation extraction," 2020, *arXiv:2010.16275*.

[66] C. Ru, J. Tang, S. Xie, S. Li, and T. Wang, "Reducing wrong labels in distant supervision for relation extraction," *Guofang Keji Daxue Xuebao/J. Nat. Univ. Defense Technol.*, vol. 40, no. 3, pp. 721–729, 2018.
[67] A. Smirnova and P. Cudré-Mauroux, "Relation extraction using distant supervision: A survey," Tech. Rep., pp. 1–35, 2019.
[68] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2007, pp. 68–74.
[69] F. Wu and D. S. Weld, "Open information extraction using Wikipedia," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2010, pp. 118–127.
[70] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, "Open language learning for information extraction," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, 2012, pp. 523–534.
[71] L. Del Corro and R. Gemulla, "Clausie: Clause-based open information extraction," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 355–366.
[72] G. Angeli, M. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process. Asian Fed. Natural Lang. Process. (ACL-IJCNLP)*, vol. 1, 2015, pp. 344–354.
[73] K. Gashteovski, R. Gemulla, and L. Del Corro, "MinIE: Minimizing facts in open information extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017.
[74] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "A survey on open information extraction," in *Proc. 27th Int. Conf. Comput. Linguistics (COLING)*, 2018.
[75] L. Cui, F. Wei, and M. Zhou, "Neural open information extraction," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 2, 2018.
[76] B. Bozic, J. K. Sasikumar, and T. Matthews, "KnowText: Auto-generated knowledge graphs for custom domain applications," in *Proc. 23rd Int. Conf. Inf. Integr. Web Intell.*, Nov. 2021, pp. 350–358.
[77] N. Kertkeidkachorn and R. Ichise, "An automatic knowledge graph creation framework from natural language text," *IEICE Trans. Inf. Syst.*, vol. E101.D, no. 1, pp. 90–98, 2018.
[78] A. Rossanez, J. C. dos Reis, R. D. S. Torres, and H. de Ribaupierre, "KGen: A knowledge graph generator from biomedical scientific literature," *BMC Med. Informat. Decis. Making*, vol. 20, no. S4, pp. 1–24, Dec. 2020.
[79] J. L. Martinez-Rodriguez, I. Lopez-Arevalo, and A. B. Rios-Alvarado, "OpenIE-based approach for knowledge graph construction from text," *Expert Syst. Appl.*, vol. 113, pp. 339–355, Dec. 2018.
[80] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019.
[81] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 601–610.
[82] F. Corcoglioniti, M. Rospocher, and A. P. Aprosio, "A 2-phase frame-based knowledge extraction framework," in *Proc. 31st Annu. ACM Symp. Appl. Comput.*, vols. 4–8, Apr. 2016, pp. 354–361.
[83] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013.
[84] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. Nat. Conf. Artif. Intell.*, vol. 28, no. 1, 2014
[85] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. Nat. Conf. Artif. Intell.*, vol. 3, 2015.
[86] Z. Sun, Z. H. Deng, J. Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, 2019.
[87] Z. Zhang, J. Cai, Y. Zhang, and J. Wang, "Learning hierarchy-aware knowledge graph embeddings for link prediction," in *Proc. 34th AAAI Conf. Artif. Intell.*, vol. 34, no. 3, 2020, pp. 3065–3072.
[88] B. Yang, W. T. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015.
[89] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *Proc. 30th AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016.
[90] T. Trouillon, J. Welbl, S. Riedel, E. Ciaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, vol. 5, 2016, pp. 2071–2080.
[91] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D knowledge graph embeddings," in *Proc. 32nd AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
[92] W. Gao and J. Wu, "Multi-relational graph convolution network for service recommendation in mashup development," *Appl. Sci.*, vol. 12, no. 2, p. 924, Jan. 2022.
[93] I. Balažević, C. Allen, and T. Hospedales, "MuReP: Multi-relational Poincare graph embeddings," 2019, *arXiv:1905.09791*.
[94] I. Balažević, C. Allen, and T. M. Hospedales, "Tucker: Tensor factorization for knowledge graph completion," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019.
[95] L. Chao, J. He, T. Wang, and W. Chu, "PairRE: Knowledge graph embeddings via paired relation vectors," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process. (ACL-IJCNLP)*, 2021.
[96] M. Wang, L. Qiu, and X. Wang, "A survey on knowledge graph embeddings for link prediction," *Symmetry*, vol. 13, no. 3, p. 485, Mar. 2021.
[97] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, and Q. He, "A survey on knowledge graph-based recommender systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3549–3568, Aug. 2022.
[98] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, and Q. He, "A survey on knowledge graph-based recommender systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3549–3568, Aug. 2022.
[99] N. Chatterjee, N. Kaushik, and B. Bansal, "Inter-subdomain relation extraction for agriculture domain," *IETE Tech. Rev.*, vol. 36, no. 2, pp. 157–163, Mar. 2019.
[100] S. Johnny and S. J. Nirmala, "Key phrase extraction system for agricultural documents," in *Proc. Int. Conf. Inf., Commun. Comput. Technol.*, in Communications in Computer and Information Science, vol. 1025, 2019, pp. 240–252.
[101] A. Nautiyal and D. Gupta, "KCC QA latent semantic representation using deep learning & hierarchical semantic cluster inferential framework," *Proc. Comput. Sci.*, vol. 171, pp. 263–272, Jan. 2020.
[102] V. Gangadharan and D. Gupta, "Recognizing named entities in agriculture documents using LDA based topic modelling techniques," *Proc. Comput. Sci.*, vol. 171, pp. 1337–1345, Jan. 2020.
[103] V. G, D. Gupta, and V. Kanjirangat, "Semi supervised approach for relation extraction in agriculture documents," in *Proc. Int. Conf. Inf. Technol. (OCIT)*, 2023, pp. 199–204.
[104] X. Guo, S. Lu, Z. Tang, Z. Bai, L. Diao, H. Zhou, and L. Li, "CG-ANER: Enhanced contextual embeddings and glyph features-based agricultural named entity recognition," *Comput. Electron. Agricult.*, vol. 194, Mar. 2022, Art. no. 106776.
[105] H. Du, V. Dimitrova, D. Magee, R. Stirling, G. Curioni, H. Reeves, B. Clarke, and A. Cohn, "An ontology of soil properties and processes," in *Proc. Int. Semantic Web Conf.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 9982, 2016, pp. 30–37.
[106] P. K. Reddy, G. V. Ramaraju, and G. S. Reddy, "eSagu: A data warehouse enabled personalized agricultural advisory system," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2007, pp. 910–914.
[107] K. Ramamritham, A. Bahuman, S. Sahni, M. Baru, C. Bahuman, A. Chandran, and M. Joshi, "The aAQUA approach: Innovative Web 2.0 tools for developing countries," *IEEE Internet Comput.*, vol. 12, no. 2, pp. 62–70, Mar. 2008.
[108] G. R. W. Arnold, "Farr, D. F.; Bills, G. F.; Chamuris, G. P.; Rossman, A. Y., fungi on plants and plant products in the united States. VIII, 1252 S. The American phytopathological society (APS) press, St. Paul (Minnesota), 1989. ISBN 0-89054-099-3," *Feddes Repertorium*, vol. 101, nos. 7–8, p. 340, 1990.
[109] S. Kübler, R. McDonald, and J. Nivre, "Dependency parsing," *Synth. Lect. Hum. Lang. Technol.*, vol. 2, no. 1, pp. 138–152, 2009.
[110] S. Rajbhandari and J. Keizer, "The AGROVOC concept scheme—A walkthrough," *J. Integrative Agricult.*, vol. 11, no. 5, pp. 694–699, 2012.

[111] C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques, and J. Keizer, "The AGROVOC linked dataset," *Semantic Web*, vol. 4, no. 3, pp. 341–348, 2013.
[112] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, nos. 4–5, pp. 993–1022, 2003.
[113] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017.
[114] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using network," in *Proc. 7th Python Sci. Conf. (SciPy)*, G. Varoquaux, T. Vaught, and J. Millman, Eds. Pasadena, CA, USA, Aug. 2008, pp. 11–15.
[115] T. Ebisu and R. Ichise, "TorusE: Knowledge graph embedding on a lie group," in *Proc. 32nd AAAI Conf. Artif. Intell.*, vol. 32. no. 1, 2018.
[116] W. Zhang, B. Paudel, W. Zhang, A. Bernstein, and H. Chen, "Interaction embeddings for prediction and explanation in knowledge graphs," in *Proc. 12th ACM Int. Conf. Web Search Data Mining (WSDM)*, Jan. 2019, pp. 96–104.
[117] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "TransD," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, 2015.
[118] S. Zhang, Y. Tay, L. Yao, and Q. Liu, "Quaternion knowledge graph embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 2459–2468.
[119] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "TransR," in *Proc. Nat. Conf. Artif. Intell.*, vol. 3, 2015.
[120] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "TransH," in *Proc. Nat. Conf. Artif. Intell.*, vol. 2, 2014.
[121] R. Abboud, I. I. Ceylan, T. Lukasiewicz, and T. Salvatori, "BoxE: A box embedding model for knowledge base completion," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 9649–9661.
[122] G. Li, L. Siddharth, and J. Luo, "Embedding knowledge graph of patent metadata to measure knowledge proximity," *J. Assoc. Inf. Sci. Technol.*, vol. 74, no. 4, pp. 476–490, Apr. 2023.
[123] I. Chami, A. Wolf, D.-C. Juan, F. Sala, S. Ravi, and C. Ré, "Low-dimensional hyperbolic knowledge graph embeddings," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020.

G. VEENA received the M.Tech. degree in machine learning from Amrita Vishwa Vidyapeetham, India. She is currently pursuing the Ph.D. degree in extracting information from text data that is specific to particular domains. She holds the position of an Assistant Professor with the Department of Computer Science and Application, Amrita School of Computing, Amritapuri, Kerala, India. Her research interests include data mining, information retrieval, machine learning, natural language processing, statistical analysis, and text analytics.

DEEPA GUPTA received the Ph.D. degree in natural language processing in example-based machine translation from the Department of Mathematics and Computer Application, Indian Institute of Technology Delhi, in 2005. She was a Postdoctoral Researcher with FBKIRST (Centre for Scientific and Technological Research), Trento, Italy, for two years on the EU-funded TC-Star Project. She is currently a Professor and the Research Head of the Amrita School of Computing, Bengaluru Campus, India. She had completed three government-funded projects and a couple of company consultancy projects. Her research interests include text analytics, clinical data mining, speech processing, and other areas in natural language processing. Her research has been published in international journals, such as IEEE Access, *Information Processing and Management*, *Knowledge-Based Systems*, and *Expert Systems with Applications*, along with several peer-reviewed international conferences.



VANI KANJIRANGAT received the Ph.D. degree in NLP, which was primarily centered on integrating machine learning and NLP techniques for text plagiarism detection. She is currently a Researcher with the Natural Language Processing (NLP) Laboratory, IDSIA, Switzerland. Ongoing research works include biomedical text mining, semantic shift detection and visual summary generations using NLP techniques, temporal embeddings, transformers, and other deep learning models. Alongside, she is also working on projects aligned with application of deep learning models in financial and question answering domains.