Knowledge Graph Embedding for Topical and Entity Classification in Multi-Source Social Network Data

Abiola Akinnubi, Nitin Agarwal, Mustafa Alassad
COSMOS Research Center
University of Arkansas - Little Rock
Little Rock, Arkansas
asakinnubi@ualr.edu, nxagarwal@ualr.edu, mmalassad@ualr.edu

Jeremiah Ajiboye
University of East London
London, United Kingdom
jeremiahajiboye@gmail.com

***Abstract*—Historically, online data has provided meaningful insights for information mining, leading to the adoption of knowledge graphs for application to online data. Knowledge embedding has become an important aspect of encoding and decoding links, relationships, and predicting the ties of an entity to an existing knowledge graph. This study applied topic modeling to extract topics, entities, and themes from heterogeneous web data from different sources around the Indo-Pacific region and modeled a knowledge graph. The knowledge graph was subjected to knowledge embedding by applying four scoring mechanisms: ComplEx, TransE, DistMult, and HolE, on a domain knowledge graph of Indo-Pacific Belt and Road initiatives to determine whether it was capable of revealing missing insights. This work significantly uses knowledge graphs and embedding to understand socioeconomic-related discussions online. Valuable insights were gained from the data in this research's clustering results of knowledge embedding. Important themes such as NASAKOM and BRI were identified in Cluster 0. Cluster 1 contained themes that discussed Marxist movements synonymous with Indonesia, and Cluster 2 showed themes on China's road policies, such as Asia-Pacific Economic Cooperation and Export-Import Bank China. Cluster 3 focused mainly on China's economic policies and the Philippines. Overall, this study demonstrates the usefulness of topic modeling and knowledge embedding in uncovering insights from online data and has implications for understanding socioeconomic trends in the Indo-Pacific region.**

***Index Terms*—Multi-Source Knowledge Graph, Knowledge Embedding, Belt and Road Initiative, Knowledge Graph, Clustering, Classification, Blogosphere**

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014- 16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21- 1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16- 1-0189, W911NF-23-1-0011), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
http://dx.doi.org/10.1145/3625007.3627315

I. Introduction

Topic modeling provides insights into the themes of discussions when applied to text. The strength of topic modeling over keywords is that it allows mapping of themes to a corpus, which is crucial in understanding documents with unlimited characters, such as blog posts. Various data modeling techniques, such as Resource Description Framework Schema (RDF Schema) and semantic networks, have traditionally inspired knowledge graphs. Historically, knowledge graph researchers have used various parts of speech to map subjects to objects. While this approach is still in use, it usually adds more noise to the data when analyzing key information. For example, many social network analysis frameworks have adopted topic modeling because it provides a pointer to what a document is about. One such study that adopts topic modeling in social network analysis is [13], in which topic modeling was used for the relational analysis of selected users' political tweet knowledge graphs. In addition, [18] used topic models to develop a knowledge graph for learning resources using approximately 500 data points extracted from various textbooks and study materials.

Unlike many approaches that use parts of speech other than entities to model relationships, particularly the predicate (relation) for knowledge graphs, topic modeling allows for sophisticated analysis of important themes and understanding of the corpus context. Knowledge graphs are powerful because a lot of information can be encoded as properties or relationships between two nodes. These relationships are called edges in traditional graphs but are referred to as predicates in knowledge graphs. Hence, clustering and classifying a graph or subgraph is always difficult, but knowledge embedding makes it possible to perform such a task. Knowledge embedding provides the opportunity to perform linear and nonlinear machine learning [1]. Knowledge graph embedding has become an important part of knowledge graph research. However, much existing knowledge embedding research has not considered long text data such as blogs, which usually contain many important themes and data from multiple social media sources.

Additionally, no work has been done to study the application of knowledge graphs to heterogeneous research, such as situational awareness of events around a region, using data from online mediums such as blogs, Twitter, and Reddit. With this in mind, an attempt has been made to model knowledge graphs from long text media such as blogs and other platforms,

namely Twitter and Reddit. Topic modeling and entity extraction were used to model relationships connecting themes, entities, and topics. Then, knowledge embedding of the graph was carried out to arrive at different metrics and interpretations using the Indo-Pacific Belt and Road Initiative as an example for data collection.

The rest of this article is organized as follows. Section II reviews existing studies to establish the foundation for this work. Section III discusses the study method and provides readers with insights into how the study was designed and how the findings were determined. Section IV discusses the results generated from the dataset used and evaluates the current efforts of this research. Section V offers insight into our future research direction and highlights the drawbacks of our current approach.

II. Related Studies

This section highlights important contributions from existing literature and how they have immensely contributed to this work. The rest of this section is divided into three subsections, which discuss the following:

1) Relation Extraction: This subsection discusses the literature that has attempted relation extraction, particularly for datasets specific to their domains.
2) Knowledge Embedding: This focuses on discussing existing literature using knowledge embedding in their research and how they used it.
3) Multi-source Social Network Analysis: Literature that used multiple datasets from different sources and knowledge graph embedding for social network analysis were considered.

A. Relation Extraction

Knowledge graphs are modeled using relationships between entities, which are crucial in constructing a knowledge graph. Relationships have existed in databases for many years; however, in natural language processing (NLP), relationships are adopted as the semantic link between a subject and an object. Scientists can now extract semantic relationships in a corpus, but constructing domain knowledge graphs remains an ongoing research domain due to data heterogeneity. The work of [2] developed an extraction method by measuring upper and lower relationships from structured data. Their work used a classification system to label web pages and applied a convolutional residual network to classify the data. The classification label was used to characterize the data as a relationship for the food domain knowledge graph. The upper and lower bounds used co-occurrence analysis to determine the implicit relationship. These extracted relationships are then used to connect the type of food to another food type.

The works of [3] constructed a knowledge graph for the legal domain by extracting relations using an improved cross-entropy loss function and bidirectional gated recurrent unit (Bi-GRU) network to extract the relationship from an unstructured legal document to enable easy case classification. The authors in [4] extracted the relationship for biomedical data and developed a knowledge graph recommendation system for biomedical data. They classified relationships between biomedical entities through their K-BiOnt system in [4], which uses knowledge graph base recommendations to improve relation classification. RelExt was developed by [5] using deep learning on the cybersecurity dataset. The approach proposed by [5] improved the cybersecurity knowledge graph, particularly for usage by security operation analysts. Their work used a semantic triple containing two cybersecurity entities to create a relation for a knowledge graph and trained the set of semantic triples using deep learning for easy relation extraction in the cybersecurity domain. The cybersecurity entities were extracted using NER (named entity recognizer) from text using a cyber-twitter system. Their approach provided three relationship classes for malware-related text, which are 'hasProduct', 'hasVulnerability' and 'uses'. The main contribution of their work is achieving a 96.61% accuracy score for the cybersecurity knowledge graph.

B. Knowledge Embedding

The authors in [10] proposed a Latent Dirichlet Allocation (LDA) augmented knowledge graph using the extracted topics from LDA to query WikiData by extracting entities from WikiData associated with the topic extracted from LDA. LDA output was used for property selection for the unstructured data because their work used structured and unstructured text. WikiData provided property descriptions. The authors in [10] used LDA to retrieve the property from WikiData and link it to the entities. The improved property selection using topics helped achieve 85% accuracy and 67% F1 score. In addition, the authors in [12] infused knowledge embedding in topic modeling tasks. Their work improved the semantic coherence of the selected topic using the TransE knowledge embedding scoring mode. The authors in [14] incorporated entity type embedding into their developed knowledge graph embedding framework called the TransET model. All entities, relations, and entity types were converted into entities in vector space. The TransET model developed by [14] achieved improved performance on existing datasets that had previously been benchmarked. Their work achieved 82.4% compared with TransE and other existing models on the same dataset. The authors in [19] presented a social network analysis-based methodology for detecting commenter mobs on YouTube. They created graph embeddings that captured the essential information of the co-commenter networks, enabling the detection of commenter mobs.

Knowledge embedding also has applicability in social-political analysis. The work of [13] used knowledge embedding on the Twitter dataset to analyze political data and classify the data. They used knowledge credibility to filter out spammers and build a knowledge graph supporting political domain facts without the noise of spammers who hijack tweets. Some of the features defined by the authors in [13] are tweet similarity, url similarity, domain content user score, domain frequency, and weight of users. Their work extracted the category of interest of selected users and generated measures

and ranking of the user's interest. Each of the extracted Twitter users, and the metadata of their interest were then used to develop a knowledge graph for each domain. The adoption of domain knowledge graphs compared with generalized knowledge graphs was due to their richness in information. The trained knowledge embedding allowed for predicting false and true political facts. Their work also used a density-based clustering algorithm (DBSCAN) for clustering with 5 clusters specified to validate their hypothesis. They were able to predict political statement associations using knowledge embedding.

C. Multisource and Social Network Analysis

Using a knowledge graph on social network data becomes natural because social network communications can be modeled from a source to a target, and the frequency of events or interactions between the source and the target becomes the weight of the graph. This is referred to as a weighted graph in classical computer science. The authors in [13] used knowledge graph embedding on Twitter datasets to study political fact relationships by first creating the modeled graph, using TransE to model graph embedding to predict the link between a node, and applying clustering and classification on the dataset.

The work of [15] extended the traditional capability of [1] to support social network analysis by adding the capability to predict ties for links and nodes. Their proposal included taking inputs $\{Actors, Ties, Ground-TruthEntities\}$ and output predicted entities. For node prediction, RLVECO allowed predicting labels for missing nodes in social graphs. The authors in [16] used hotspot data over 10 years and developed a knowledge graph using the frequency of related keywords and centrality measures to analyze medical data. The work of [17] also used frequency and weight occurrence to model tweet popularity and combined knowledge embedding techniques to better represent Twitter data. Their representation achieved an improved error rate of -5% and +17% hit rates compared with the state-of-the-art benchmark. Leveraging a co-occurrence graph, their approach used the adjacent matrix for a learning graph convolution network (GCN). They used GCN on a subgraph of each co-occurrence graph and then used weighted embedding based on the frequency of the output generated by the GCN. The use of knowledge graphs and knowledge embedding continues to gain attention in traditional social network analysis because relationships are modeled as graphs.

III. Methodology

This section describes the data used in this study, the approach used in modeling the relationship between entities, and the topic modeling of topic words extracted from the corpus. Figure 1 shows the high-level overview of the methodology used in this study. This section is divided into four subsections:

1) Dataset subsection describes how the data used in this study were collected.
2) Relation modeling subsection describes how the knowledge graph developed from the collected data was modeled and stored in the neo4j database.
3) Knowledge graph embedding subsection discusses how knowledge embedding from the graph stored in the neo4j database was achieved and how the data on various algorithms were trained.
4) Embedding and Classification subsection discusses how knowledge embedding and classification tasks were achieved.

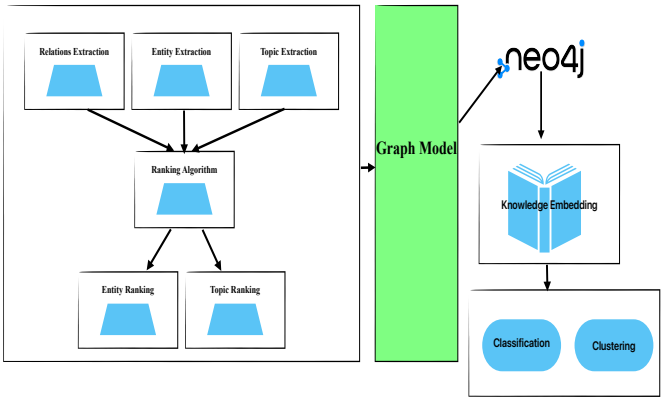

Fig. 1. Illustration of the overview of our entire approach.

A. Dataset

Data collection efforts focused on the Belt and Road Initiative issue, specifically targeted at Indonesia. Our data contains blog posts, Reddit posts, and tweets published between January 2019 and November 2022. We matched this data with a key phrase that may exist in the contents of the collected data. The italicized quoted text below shows some of the key phrases used in the study. For this experiment, we focused on a sample size of 300,000 datasets. We decided it was best to extract entities and topics from longer, continuous sentences rather than short sentences for the generated knowledge graph to have rich node counts.

'antek', 'aseng asing', 'Tiongkok', 'Tionghoa Cina China', 'Indonesia', 'Cina', 'OBOR', 'BRI', 'kebijakan', 'luar', 'negeri', 'proyek', 'pekerja', 'OBOR BRI', 'Cina', 'China', 'Tionghoa', 'Tiongkok', 'Pembangkit Listrik Tenaga Batubara', 'Cina China', 'Perusahaan Listrik Negara', 'BRI OBOR', 'Proyek 35000 Megawatt', 'Maritime Silk Road', 'Jakarta Indonesia', 'Global Maritime Fulcrum', 'Jokowi', 'Tiongkok China Cina Tionghoa OBOR BRI', 'Menguasai', 'Cina China Tiongkok Tionghoa', 'ekonomi', 'Pekerja Cina pulang', 'Chinese workers go home'.

B. Relation Modeling

To develop a knowledge graph that generates a binary relationship is important. These relationships could be semantic text or induced association, i.e., association generated from prehistorical knowledge of the corpus context and newly mined information. Many researchers use open source Knowledge Graph already developed for their domain of interest. For example, in the works of the author in [4], we believe that

this approach lacks scalability particularly when experiments require the use of heterogeneous data from multiple web sources. Some existing data are limited to the *New York Times* or any specific web domain. Therefore, this study collects data from multiple sources discussing the Belt and Road Initiative to explore how we can model relationships and translate the relationships into a knowledge graph that can be used for training knowledge embedding.

As part of this study's ongoing efforts to model relationships for a multi-source knowledge graph, the approach involved extracting entities' named entity relation (NER) from a multi-source corpus, since the multi-source corpus provides rich text content to extract multiple useful entities for this purpose. Topics were also extracted using Gensim LDA for topic modeling. To control the number of nodes for our knowledge graphs, the focus was limited to 5 types of entities, PERSON, ORG, NORP, GPE, LOC, and EVENT, while 10 topics were used for the Topic nodes. Each topic contains some topics or themes discussed by the corpus content extracted from the text using the Genism LDA library. Table I shows the definition of each NER used in this study. Furthermore, the topic words were collected from the LDA,

TABLE I
SELECTED TYPE OF EXTRACTED ENTITIES AND THEIR MEANING

NER Type	Meaning
PERSON	A particular individual e.g Abiola, Trump
ORG	Organization
LOC	Location
EVENT	Event
NORP	NORP

and each was intended to be used as a node for the study's graph; these resulted in having three important nodes for this purpose which are Entity, TopicWord, and TopicNumber, while the relation is defined between a TopicWord (TWord) and a TopicNumber as 'TOPIC_CONTAINS_TWORD'. The relationship between entities and topics is defined as 'TWORD_IS_MEMBER_OF_TOPIC'. Using knowledge graph embedding, we modeled this study's graph to test whether it is possible to predict the likelihood that any word collected from the same region would belong to the topic or have a link to a topic, entity, and topic word. Hence, we decided to first, as shown in the Algorithm 1 below, generate and extract topic models, entity extraction, and topic words from the text corpus. The first step was to iterate through the blog data frame. The next step was to add an extra attribute to the data, which is stored as a dictionary containing the topic mapping and which topic is the most ranked using the topic with the highest probability scores. That is, if a topic has the highest score, it becomes the label of a data document. A document's highest probability could be topic 1, 2, or 3. The words that belong to the most rated topics for each computation for each document were also mapped.

Algorithm 1: Corpus enhancement with topic and entity

Definitions: 1
D_1,D_2,D_n represents each document in multisource document.
Dn_t represents the computed topics for the current document
Dn_e represents the computed entities for the current document
Dn_k represents the computed keywords for the current document
Inputs : $dataset \leftarrow \{D_1,D_2,D_n\}$
Output : $dataset$ which is modified and enhanced
EnhanceDataset $(dataset)$
  foreach *corpus belonging to* $D_n \in dataset$ **do**
    $Dn_e \leftarrow$ **ExtractEntity**$(corpus)$
    $Dn_t \leftarrow$ **ExtractTopic**$(corpus)$
    $Dn_k \leftarrow$ **ExtractKeyword**$(corpus)$
  return $dataset$

After enhancing the data with the extracted entities and topics, the required triple type was generated, which is crucial for developing knowledge embedding. This involves modeling the data in the format D = $(subject, predicate, object)$, where the subject, in this case, is the entity and the predicate is the relationship termed as 'TWORD IS MEMBER OF TOPIC' which shows that an entity is related to a topic. This can be invariably described as G = { (ENTITY, BELONGS TO, TOPIC NUMBER)} and G = {(TOPIC WORD, TOPIC WORD IS CONTAINED IN, TOPIC NUMBER)} or put G = $(node, relationship, node) \subseteq$ node X relationship X node is a set of $(node, relationship, node)$ triples, each including a subject sub $\in$ node, a predicate relationship $\in$ R, and an object node $\in$ node. $node$ and R are the sets of all entities and relation types of G. Hence, the entity as the source node in Algorithm 2 was used with the associative statement as a relation and the target topic number as the target node. The same procedure was repeated for the topic words. An associative statement to show the relation of a topic word's presence in a topic number node Algorithm 2 was used, as shown in Figure 2. This allows a data frame of the subject-relation-object to be returned. This is then used to train the knowledge embedding algorithm described in Section III-C.

C. Knowledge Graph Embedding

This section discusses the approach used for knowledge graph embedding. We stored the constructed knowledge graph from the extracted and modeled data in Section III-A in the Neo4J property graph database. We then retrieved the stored data using cypher query, which was retrieved as (subject, predicate, object). We then split the retrieved collection of (subject, predicate, object), which is represented as $\{\ G = \{(s,p,o)_1, (s,p,o)_n\}\ \}$ and which is then split into training and test sets. A total of 10,000 nodes were

Algorithm 2: Topic-entity-word triple model

Inputs : $dataset \leftarrow \{D_1, D_2, D_n\}$
Output : $tripleList$
GenerateTriple $(dataset)$
  $tripleList \leftarrow$ **list**
  foreach $data$ *in* $dataset$ **do**
    $topicLabels \leftarrow data.\textbf{get}('topicLabel')$
    $entites \leftarrow data.\textbf{get}('entity')$
    $topicWords \leftarrow data.\textbf{get}('topicWord')$
    $topicEntityMappingList =$
    $[topicLabels, entities, topicWords]$
    foreach $topicEntity \in$
     $topicEntityMappingList$ **do**
      foreach $entity \in topicEntity$ **do**
        $triplePairs \leftarrow [entity.\textbf{get}(name),$
        $topicEntity.\textbf{get}(topic),$
        $topicEntity.\textbf{get}(word)]$
        $tripleList.\textbf{append}(triplePairs)$
  return $tripleList$

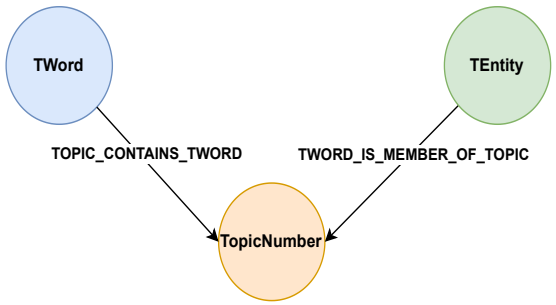

Fig. 2. A visualization showing entity-topic-word graph model.

used for training the data from the retrieved 94,415 with a total of 104,650 existing between the nodes. Relationships are of two categories defined by us for this research purpose, namely: 'TOPIC_CONTAINS_WORD' and 'IS_MEMBER_OF_TOPIC'.

We then used Ampligraph [1] to train the dataset. Several scoring models have been supported for knowledge graph embedding tasks; some are, with their benchmarks against publicly available datasets, shown in the table below. This serves as a foundation to evaluate if a custom-modeled knowledge graph can achieve a higher score for various knowledge embedding scoring models with the top 1, 3, and 10 hits, respectively, and with the MR scores. These datasets are standardized literature. We aim to predict if a link exists between any given entity that is defined to be from the Belt and Road Initiative data that the knowledge embedding model has not seen before and evaluate the performance. The framework used for performing this task is shown in the visualization Figure 3. The graph used in training the model consists of two groups: the word-topic knowledge graph and the entity-topic knowledge graph.

We used the ComplEx scoring type in this study with an Adam optimizer (an Adam optimizer is defined as an optimizer that estimates the first and second gradient to adapt its learning rate for each weight of the graph learning). We configured our work to use a batch size of 10,000 and an epoch of 300. We chose 300 as the epoch to find a balance to avoid overfitting with the sizeable number of relationship categories modeled in the knowledge graph we developed. We discuss the evaluation results in the results and discussion section. We then used the trained and saved model to achieve an embedding clustering task. We will discuss our clustering and classification method next.

D. Embedding Clustering and Classification

To determine how the best knowledge graph and knowledge embedding can benefit other aspects of social network analysis, we applied clustering and classification methods using the output generated from section III-C. Since traditionally, graphs are measured by weights, centrality measures, and the likes, other forms of social analysis usually involve classification or clustering tasks. This is because knowledge embedding helps simplify graphs to allow classification. We used the embedded graph results from Section III-C to generate topic word and k-mean clustering for entity clusters on knowledge graph embedding data. We specifically used a cluster size of five in this study, which was obtained by finding the best cluster using the silhouette function to identify the optimal cluster point. We wanted to see if similar topic words and entities would appear close to each other. This solidified why using clustering could help test this hypothesis.

Furthermore, we adopted a BERT-based classification model trained using the relationship link output generated by knowledge graph embedding to label our dataset. Then, we trained the model to see if we could use the classification task on the output generated by knowledge graph embedding. We used the cross-entropy loss function for the bidirectional encoder representations from transformers (BERT) classifier with the ReLU (rectified linear unit) activation function to train our BERT classifier model. We also used the Adam optimizer that was defined earlier in the knowledge graph embedding section III-C, and set the learning rate to 5e-5. In this work, we also decided to use an epoch of 300 for this model to classify our knowledge graph embedding output. We stored this model and attempted to use it to predict and classify both topic-word and topic-entity relationships.

IV. Discussion

A. Knowledge Graph Construction

As discussed in our methodology, our use case contained multi-source data collected for Belt and Road initiatives (BRI). We modeled our knowledge graph with nodes containing information extracted from topic modeling and entity extractions. We observed that the influential nodes for topics discussed in the text corpus are China, Indonesia, and Orange, as shown in Figure 4. These themes aligned with the key phrases used in our data collection, as stated in Section III-A. This proves

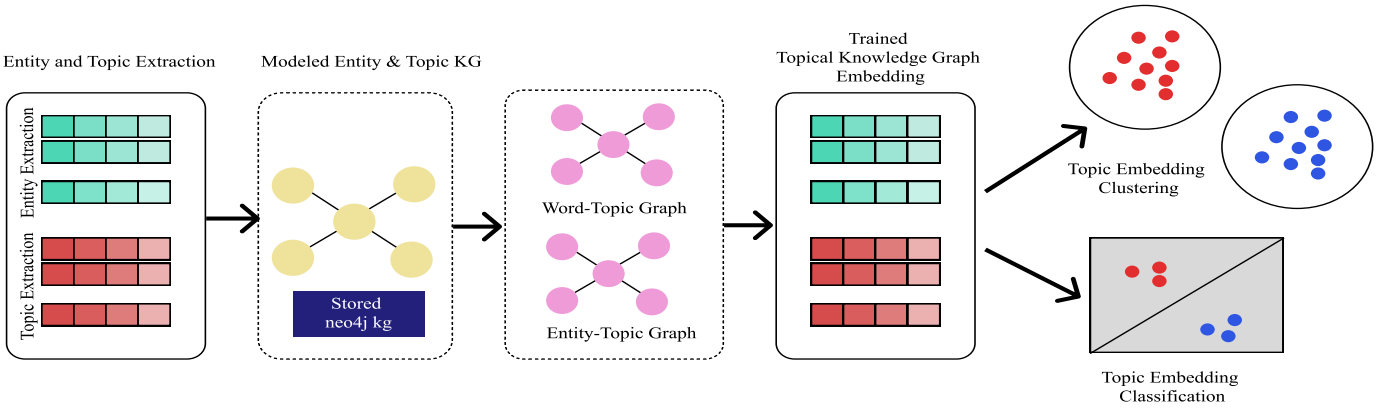

Fig. 3. A visualization showing entity-topic extraction to Knowledge graph data and generated trained knowledge embedding for Belt and Road Initiative Indo-Pacific data with clustering and classification output

TABLE II
SELECTED BENCHMARKED MODEL AVAILABLE VIA [1] USING AMPLIGRAPH

Benchmarked Data	Model	MR	MRR	Hits@1	Hits@3	Hits@10
FB15K-237	TransE	211	0.31	0.22	0.34	0.48
	DistMult	211	0.30	0.21	0.33	0.48
	ComplEx	197	0.31	0.21	0.34	0.49
	HolE	190	0.30	0.21	0.33	0.48
YAGO3-10	TransE	1210	0.50	0.41	0.56	0.67
	DistMult	2301	0.48	0.39	0.53	0.64
	ComplEx	3153	0.49	0.40	0.54	0.65
	HolE	7525	0.47	0.38	0.52	0.62
WN18RR	TransE	3143	0.22	0.03	0.38	0.52
	DistMult	4832	0.47	0.43	0.48	0.54
	ComplEx	4229	0.50	0.47	0.52	0.58
	HolE	7072	0.47	0.44	0.49	0.54
FB15K	TransE	45	0.62	0.48	0.72	0.84
	DistMult	227	0.71	0.66	0.75	0.80
	ComplEx	199	0.73	0.67	0.77	0.82
	HolE	222	0.72	0.65	0.77	0.83
WN18	TransE	278	0.64	0.39	0.87	0.95
	DistMult	729	0.82	0.72	0.92	0.95
	ComplEx	758	0.94	0.94	0.95	0.95
	HolE	676	0.94	0.93	0.94	0.95

that using topic extraction to construct knowledge graphs is important, particularly when the data is heterogeneous and unlabeled, or not annotated. Figure 4 shows entities related to the topic Indonesia, and our knowledge graph shows that the Indonesia node is related to or belongs to topic 3 of the 10 topics extracted from the dataset.

Entities closely associated with the Indo-Pacific region have entity mentions like ‘Gowak’ and ’Bong Swi Hoo Gowak,’. Our knowledge graph of the topic-entity relationship is shown in Figure 5. We have shown four different topics and entities connected to these topic themes. In Figure 5, we observed nodes such as ’China’, protest’, ’Cheng Ho’, and others for the entities-topic relationship under topic 3.

B. Knowledge Graph Embedding Model

After storing the knowledge graph in the Neo4j graph database, [1] is leveraged to perform knowledge embedding

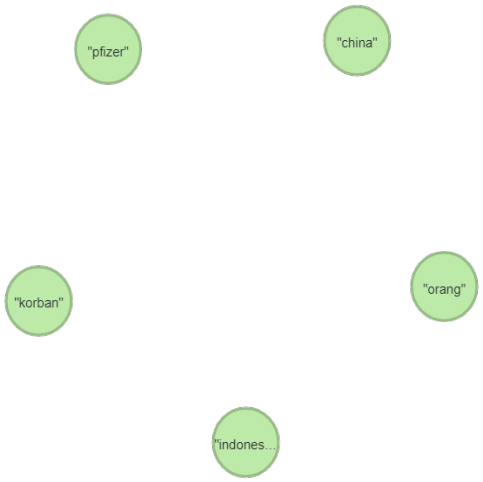

Fig. 4. Topic theme node of documents for Belts and Road Initiatives

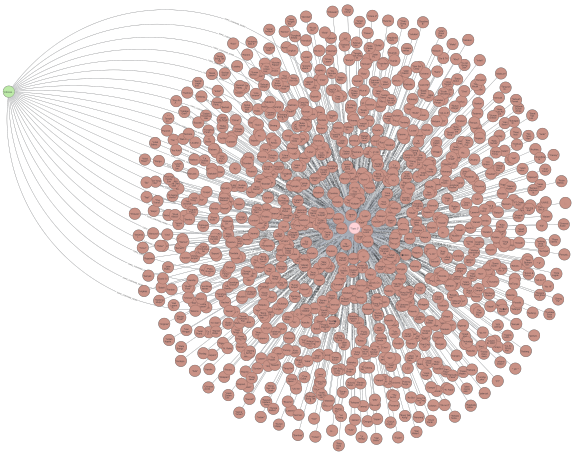

Fig. 5. Indonesia node and connected entity from Neo4j knowledge graph constructed from topic and entities

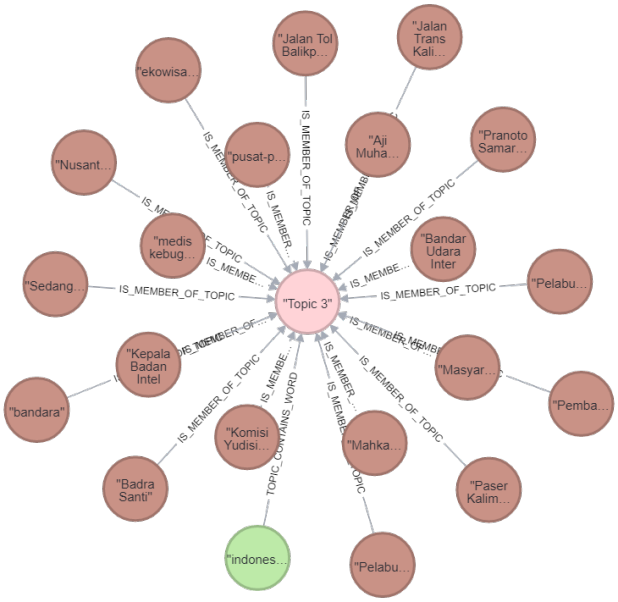

Fig. 6. Knowledge graph diagram for topic 3 showing a direct relationship between the topic and entities

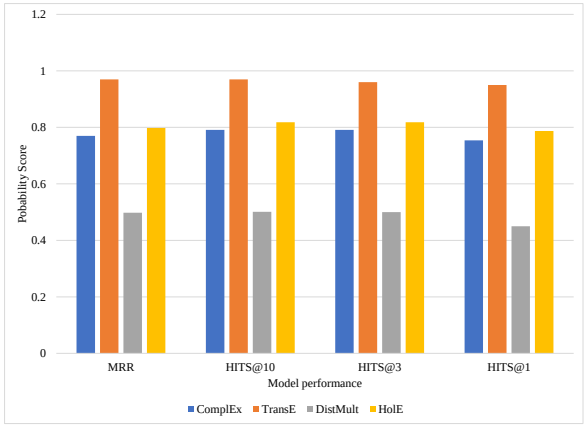

Fig. 7. Histogram chart comparison of model performance

because the domain data are in the knowledge graph format. A comparison of the model performance on our extracted graph with other benchmark data was attempted to determine how the model would perform for the specified task. Scores were computed for different types of graph embedding models, although TransE has been a very popular graph embedding scoring model widely accepted in the community. The four models used in this study were ComplEX, TransE, DistMult, and HolE. An MRR (Mean reciprocal rank) score was generated to determine how closely this study's knowledge graph embedding model would predict the link for a given task in the Indo-Pacific region. The results show that ComplEX, TransE, and HolE predict the link for a given task with a probability of more than 70%, although this study's TransE model performed better on this dataset. Table III and Figure 7 show the side-by-side model performance.

TABLE III
SCORING RESULTS FOR FOUR MODELS TRAINED ON THIS STUDY'S KNOWLEDGE GRAPH. NOTES HIT STATE HOW THIS STUDY'S MODELS WILL PERFORM WHEN THE TOP 10, 3, AND 1 ARE SELECTED

Model	Epochs	MRR	HITS@10	HITS@3	HITS@1
ComplEx	300	0.77	0.791	0.791	0.754
TransE	300	0.97	0.97	0.96	0.95
DistMult	300	0.498	0.501	0.50	0.45
HolE	300	0.798	0.818	0.818	0.787

It was observed that this model performed well with the defined relationship of a graph because the data were carefully modeled to reflect the themes and entities mentioned in the collected data. This model benefited from relationships that connect topics, topic number, and entities stored as graphs. It was observed that this study's knowledge graphs had a higher probability MRR (mean rank) for translational models (models that apply linear transformation in their scoring) because they have a pairwise loss function.

C. Knowledge Graph-Embedding Entity Clustering Analysis

Finally, the generated knowledge graph embedding was used to perform clustering, and k-mean clustering was performed by selecting only four clusters for our data as shown in Figure 8. Knowledge embedding allows performing machine learning tasks such as clustering. For all clusters, it was observed that Cluster 0 contained words that discussed ideology around democratic movements in Indonesia. For example, NASAKOM is a socialist movement in Indonesia and China roles; it also had words mentioning China. Cluster 1 also mentioned Marxist individuals and movements, a form of social and political ideology (Muhammad al-fayyadl, Union Soviet Socialist Republics). In contrast, Cluster 2 contains words with themes centered on China's economic expansion and its economic role in the Asian region. Table IV contains some important themes that could be extracted through the knowledge embedding results when clustering is applied. Finally, this study's approach and use case has further shown how much of hidden information in a multisource heterogeneous data can be mined when knowledge graphs, embedding, and clustering are applied. Modeling this information in knowledge graphs allowed us to gain insights that would have otherwise been unknown.

V. FUTURE WORKS

This current approach will require improving the labeling of relationships for heterogeneous data from the extracted information to better understand the stances of online actors or entities. This will improve the prediction of links between influential actors and the audience with which they are targeted with factual information or disinformation. While work on this is in progress, it is important to highlight that future work will

TABLE IV
SOME WORDS FROM THE SELECTED CLUSTERING RESULTS OF KNOWLEDGE EMBEDDING

cluster 0	cluster 1	cluster 2	cluster 3
BRI	Covid Indonesia	21st Century Maritime Silk Road 2013	Armed Forces Philippines
NASAKOM	Muhammad Al-Fayyadl	Asia-Pacific Economic Cooperation	Asia Infrastructure Investment Bank
DPRD DKI	Munculnya Reynhard Sinaga	Export-Import Bank China	Asian Financial Crisis
People of Republic China	Union Soviet Socialist Republics	International Monetary Fund	BRI China
Army Indonesian Genocide Mechanics Mass Murder of Jess Melvin	Ruling Tribunal UNCLOS 2016	One Belt One Road	Quadrilateral Security Dialogue

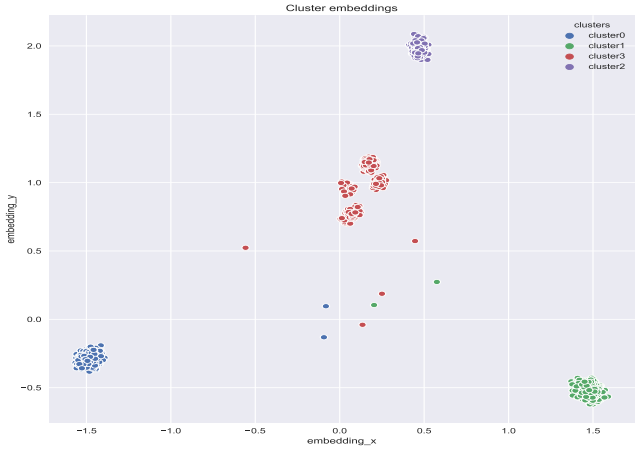

Fig. 8. Clustering plot generated from knowledge graph embedding

consider using knowledge graph embedding for morality and emotion analysis assessments.

REFERENCES

[1] Luca Costabello and Sumit Pai and Chan Le Van and Rory McGrath and Nick McCarthy and Pedro Tabacof, AmpliGraph: a Library for Representation Learning on Knowledge Graphs, 2019, https://doi.org/10.5281/zenodo.2595043

[2] H. Yu, H. Li, D. Mao, and Q. Cai, "A relationship extraction method for domain knowledge graph construction," WORLD WIDE WEB-INTERNET AND WEB INFORMATION SYSTEMS, vol. 23, no. 2, SI. SPRINGER, ONE NEW YORK PLAZA, SUITE 4600, NEW YORK, NY, UNITED STATES, pp. 735–753, Mar. 2020. doi: 10.1007/s11280-019-00765-y.

[3] B. Dong, H. Yu, and H. Li, "A Knowledge Graph Construction Approach for Legal Domain," TEHNICKI VJESNIK-TECHNICAL GAZETTE, vol. 28, no. 2. UNIV OSIJEK, TECH FAC, TRG IVANE BRLIC-MAZURANIC 2, SLAVONSKI BROD, HR-35000, CROATIA, pp. 357–362, Apr. 2021. doi: 10.17559/TV-20201119084338.

[4] D. Sousa and F. M. Couto, "Biomedical Relation Extraction With Knowledge Graph-Based Recommendations," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 8, pp. 4207–4217, Aug. 2022, doi: 10.1109/JBHI.2022.3173558.

[5] A. Pingle, A. Piplai, S. Mittal, A. Joshi, J. Holt, and R. Zak, "RelExt: relation extraction using deep learning approaches for cybersecurity knowledge graph improvement," in Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, in ASONAM '19. New York, NY, USA: Association for Computing Machinery, Aug. 2019, pp. 879–886. doi: 10.1145/3341161.3343519.

[6] C. Zheng, "Comparisons of the City Brand Influence of Global Cities: Word-Embedding Based Semantic Mining and Clustering Analysis on the Big Data of GDELT Global News Knowledge Graph," Sustainability, vol. 12, no. 16, p. 6294, Aug. 2020, doi: 10.3390/su12166294.

[7] F. Al-Obeidat, O. Adedugbe, A. B. Hani, E. Benkhelifa, and M. Majdalawieh, "Cone-KG: A Semantic Knowledge Graph with News Content and Social Context for Studying Covid-19 News Articles on Social Media," 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 1–7, Dec. 2020, doi: 10.1109/SNAMS52053.2020.9336541.

[8] T. A. Munna and R. Delhibabu, "Cross-Domain Co-Author Recommendation Based on Knowledge Graph Clustering," N. T. Nguyen, S. Chittayasothorn, D. Niyato, and B. Trawiński, Eds., in Lecture Notes in Computer Science, vol. 12672. Cham: Springer International Publishing, 2021, pp. 782–795.

[9] C. Li, X. Chen, Y. Zhang, S. Chen, D. Lv, and Y. Wang, "Dual Graph Embedding for Object-Tag Link Prediction on the Knowledge Graph," 2020 IEEE International Conference on Knowledge Graph (ICKG), pp. 283–290, Aug. 2020, doi: 10.1109/ICBK50248.2020.00048.

[10] P. B. Abels, Z. Ahmadi, S. Burkhardt, B. Schiller, I. Gurevych, and S. Kramer, "Focusing Knowledge-based Graph Argument Mining via Topic Modeling," ArXiv, Feb. 2021, Accessed: Jun. 01, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Focusing-Knowledge-based-Graph-Argument-Mining-via-Abels-Ahmadi/bd429d49ac29aa8ba9c2267905657ac7aaacfe39

[11] M. Brambilla and B. Altinel, "Improving Topic Modeling for Textual Content with Knowledge Graph Embeddings," presented at the AAAI Spring Symposium Combining Machine Learning with Knowledge Engineering, 2019. Accessed: Jun. 01, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Improving-Topic-Modeling-for-Textual-Content-with-Brambilla-Altinel/ab3e352affeceabc35bab1b9628d5a2f6443acf2

[12] L. Yao et al., "Incorporating Knowledge Graph Embeddings into Topic Modeling," in Proceedings of the AAAI Conference on Artificial Intelligence, Feb. 2017. doi: 10.1609/aaai.v31i1.10951.

[13] B. Abu-Salih et al., "Relational Learning Analysis of Social Politics using Knowledge Graph Embedding," Data Min Knowl Disc, vol. 35, no. 4, pp. 1497–1536, Jul. 2021, doi: 10.1007/s10618-021-00760-w.

[14] P. Wang, J. Zhou, Y. Liu, and X. Zhou, "TransET: Knowledge Graph Embedding with Entity Types," Electronics, vol. 10, no. 12, p. 1407, Jun. 2021, doi: 10.3390/electronics10121407.

[15] B. C. Molokwu and Z. Kobti, 'Social Network Analysis using RLVECN: Representation Learning via Knowledge-Graph Embeddings and Convolutional Neural-Network', in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, Jul. 2020, pp. 5198–5199. doi: 10.24963/ijcai.2020/739.

[16] L. Jiahui, Z. Peiyao, and Y. Xiaoliang, 'Research Hotspots and Trends Analysis of Real-World Data Based on Social Network Analysis and Knowledge Graph', Asian Journal of Social Pharmacy vol. 16, no. 3, pp. 272–279, 2021.

[17] Y. Zhang, X. S. Fang, and T. Hara, 'Evolving Social Media Background Representation with Frequency Weights and Co-Occurrence Graphs', ACM Trans. Knowl. Discov. Data, vol. 17, no. 7, pp. 1–17, Aug. 2023, doi: 10.1145/3585389.

[18] A. Badawy, J. A. Fisteus, T. M. Mahmoud, and T. Abd El-Hafeez, "Topic Extraction and Interactive Knowledge Graphs for Learning Resources," SUSTAINABILITY, vol. 14, no. 1. MDPI, ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND, Jan. 2022. doi: 10.3390/su14010226.

[19] S. Shajari, N. Agarwal, and M. Alassad, "Commenter Behavior Characterization on YouTube Channels," Apr. 2023, doi: 10.48550/ARXIV.2304.07681.