RAING: Retrieval Augmented generation Inside kNowledge Graph: An Approach to unveil Fake News

Orlando Abaunza Ubaque and Diego Rincon-Yañez

***Abstract*—The phenomenon of fake news (FN) is a growing concern due to its ability to distort public perception of reality and undermine trust in information sources. This work addresses the problem of fake news detection using a combination of Knowledge Graphs (KGs) and Large Language Models (LLMs). Current challenges in fake news detection are analyzed, and a methodology for integrating KGs into fake news detection using the Retrieval-Augmented Generation (RAG) approach is proposed. Various strategies for building and utilizing KGs in fake news detection are discussed, highlighting the importance of integrating external information to improve accuracy and contextualization. The proposed methodology is divided into key stages, including data management, model learning, model verification, and results analysis. techniques used in each stage, including data selection, KG construction, KG embedding generation, and RAG implementation, are detailed. A set of primary and secondary objectives are established to evaluate the effectiveness of the proposed approach and an evaluation methodology including metrics such as precision, compelling misinformation or hallucinations detection.**

***Index Terms*—Fake news, Knowledge Graphs, Knowledge Graph Embeddings, shortest neighbour search, Large Language Models, Retrieval-Augmented Generation**

I. Introduction

A. Problem Statement

CURRENTLY the phenomenon of fake news (FN) is a growing concern due to its ability to distort public perception of reality and undermine trust in information sources. Notable international cases include the 2016 US elections, where almost as many fake news stories as real ones were shared [1] and in Colombia, cases such as those of Andres Sepulveda, who confessed to leading a team of hackers to manipulate social media and create false waves of enthusiasm during the 2014 Colombian elections [2]. The impact of such news can be significant, as evidenced during the pandemic when, according to research [1], mortality rates were higher in countries where social media was the main source of information and where there was greater trust in its content, leading to public policies and even imprisonment for those who spread false news about COVID-19 in Peru[2].

According to the classification by Allcott and Gentzkow [3], there is disinformation (1) news created with the intention of deceiving, such as those about elections or COVID-19, and misinformation (2) news produced without the intention of deceiving but which ends up being misleading, as often happens with Large Language Models. This is one of the problems of these models, according to Dziri [4], who referred to it as "hallucinations" when they generate contexts that contradict or mislead real-world knowledge.

This issue has been addressed in various ways, depending on the context. Initially, the syntactic relationship, such as Bag of Words or n-grams, was examined to classify fake news based on syntactic similarity to verified news, neglecting the semantic relationship and the veracity of the information since it could not be contrasted against facts [5]. The emergence and diffusion of Large Language Models (LLMs) marked a milestone for detecting fake news, thanks to their capacity for semantic analysis, capturing both the structure and meaning of the text, which allows them to identify more complex patterns and discern between true and false news with greater accuracy. Despite these advantages, they fall short when it comes to validating news outside the context in which the LLMs were trained, such as information about specific companies or news released after a particular LLM was launched.

To address this shortcoming, various strategies have been developed, ranging from more aggressive approaches that modify the model by retraining it to include new information, commonly known as fine-tuning, with the underlying risk of deteriorating their overall performance and generating more hallucinations [6], in addition, it can cause the models to mechanically memorize the facts without true understanding, which limits their ability to effectively integrate new information [7], to less invasive ones that add new knowledge as an additional source of consultation for the model[8]. Particularly, Shirui [9] favors the latter,

Submitted: 22-11-2024 , **Defended:** 21-01-2025

O.A.U. was with Pontificia Universidad Javeriana Bogotá, CO. As a student in the Artificial Intelligence Master Program (e-mail: abaunzao@javeriana.edu.co).

D. RY., as Supervisor, is currently a Research Fellow in the School of Computer Science and Statistics at Trinity College Dublin, IRL (e-mail: diego-rincon@javeriana.edu.co).

Jurors: Leonardo Florez-Valencia, (Pontificia Universidad Javeriana) and Karel Gutiérrez-Batista (Universidad de Granada).

[1]Quartz: https://qz.com/1090903/people-shared-nearly-as-much-fake-news-as-real-news-on-twitter-during-the-election

[2]Bloomberg: https://www.bloomberg.com/features/2016-how-to-hack-an-election

reviewing different strategies to improve LLMs using external knowledge in the form of graphs, which are data structures that use nodes (or vertices) and edges (or connections) to represent and organize complex information, where they propose three main ways to integrate graphs into LLMs: through the pre-training of KG-enhanced LLMs, which aims to inject knowledge into LLMs during the pre-training stage; through KG-enhanced LLM inference, which enables LLMs to consider the latest knowledge while generating sentences; and finally, in the interpretability of KG-enhanced LLMs, which aims to improve the interpretability of LLMs using KGs. The methodology of KG-enhanced LLM inference is particularly attractive, as it allows for updating and exploiting LLMs at a low cost without retraining the model. It is versatile in its fields of application, as it can be used for new or specific information on a particular topic or company. Knowledge graphs (KG) have become essential for organizing and representing knowledge in various domains, supporting applications such as question answering systems (QA) [10]. The combination of OpenIE and KG embedding models, like REBEL and TransE, enables efficient fact extraction and retrieval, enhancing QA performance in specialized domains.

This work proposes an approach to detecting fake news by combining knowledge graphs, vector spaces, and language models. Begins with creating a base knowledge graph (KGBase) generated by extracting triples (subject, predicate, object) from a news dataset. From this graph, a vector space is constructed using embedding algorithms such as RotatE and TransE, enabling the numerical representation of entities and relationships. To validate a news article, it is decomposed into triples and then compared with the KGBase and its vector representation, identifying similar entities and relationships through nearest-neighbor searches. This process creates an enriched semantic context in which entities, relationships, and patterns are integrated. Finally, this information is sent to a large language model (LLM), which uses augmented knowledge to analyze the news and generate more accurate predictions, offering a robust and well-founded method to combat misinformation.

B. Motivation and Relevance

The integration of graph embeddings in knowledge, a still recent field, offers an innovative solution by enriching LLMs with external and up-to-date information. This approach allows models to handle explicit and implicit queries more efficiently, avoiding the superficial retrieval of irrelevant data and improving the quality of generated responses [9]. By integrating these types of techniques, organizations can improve the performance of their AI systems without significant investments, improving their competitiveness in the global digital environment.

II. Objectives

A. Main Objective:

Evaluate the effectiveness of RAGS in Large Language Modes (LLMs) for injecting data out of the training base and augmenting using knowledge graph embeddings in the context of fake news detection.

B. Secondary Objectives:

1) Select an open dataset allowing for meaningful comparisons between fake and real news.
2) Generate knowledge graph embeddings from the selected dataset.
3) Build a Retrieval Augmented Generation (RAG) based on the selected dataset for fake news detection.
4) Evaluate the effectiveness of the proposed approach using experiments and metrics in fake news detection.
5) Conduct rigorous experiments to validate the effectiveness of the proposed Retrieval Augmented Generation using precision.
6) Evaluate the effectiveness of the proposed Retrieval-Augmented Generation approach in fake news detection using a comprehensive set of metrics, including precision, accuracy, contextual understanding, compelling misinformation, transparency and traceability, and source retrieval accuracy.

III. Related Work

This section provides an overview of key approaches and advancements related to the detection of fake news and the challenges associated with hallucinations in Large Language Models (LLMs), which generate content that contradicts real-world knowledge (Dziri et al., 2021[4]). These issues can be mitigated through fine-tuning, but this approach incurs significant costs in terms of time and computational resources (Agrawal et al., 2023[6]). We explore four main areas of research: analysis of writing style features for identifying misinformation, the integration of knowledge graphs(KG) to enhance contextual understanding, the application of Retrieval-Augmented Generation (RAG) methods for reducing hallucinations, and the role of embeddings in improving the reasoning and accuracy of LLMs.

A. Detection of false notifications through writing style

Numerous studies have focused on detecting fake news on social media platforms like Twitter using machine learning models. These studies extract linguistic features such as n-grams from textual articles to train various machine learning models like K-nearest neighbor (KNN), support vector machine (SVM), and stochastic gradient descent (SGD). The highest accuracy (92%) was achieved with SVM and logistic regression, although according to Wang[11] as the n-gram size increases, the precision decreases. Additionally, incorporating auxiliary user information has shown better results in some studies. In the dEFEND[12] work, they also considered user information in addition to analyzing comments on the news to find patterns that could lead to verifying the validity of a particular news story MIPR[13].

These techniques are mainly based on linguistic features extracted from textual articles, such as n-grams, to classify news. However, they often struggle with the dynamic and nuanced nature of the language used in news articles, making it challenging to accurately discern between true and false information. Additionally, these methods may overlook semantic relationships and context, which are crucial for verifying the veracity of news content. As disinformation tactics evolve and adapt, traditional machine learning models may find it difficult to keep pace, leading to reduced effectiveness in identifying fake news.

B. Knowledge Graph for Fake News Detection

Knowledge graphs used for fake news detection have been notable for their ability to capture and represent complex semantic relationships between entities. Manual information extraction and annotation can be employed to generate these KG, which are fundamental techniques especially useful for smaller datasets or when precision in relationship capture is required, as seen in the work of Jaradeh et al. [14]. Meanwhile, ontologies provide a structured model for representing concepts and relationships within a specific domain, a subject explored by Krötzsch and Thost when discussing ontology rules for KG, highlighting the need to adapt existing ontologies to standardized KG formats [15]. Furthermore, open-source tools like REBEL facilitate KG construction by extracting entity triplets and their relationships. This model enables efficient hierarchical organization and is particularly useful for handling large volumes of data and automating concept classification[16]. The work of KAN:[17] "Knowledge-aware Attention Network" utilizes attention mechanisms to integrate external knowledge from KGs into text summarization. It applies attention mechanisms to measure the importance of each entity and its context, feeding this information into a neural network, thereby improving accuracy by providing a richer and more relevant context. This approach can be adapted not only to retrieve information from a text corpus but also to validate this information using a similarity function.Shirui [9] favors the latter, reviewing different strategies to improve LLMs using external knowledge in the form of graphs.

In the work Koloski[18] the integration of KG into fake news detection is proposed through heterogeneous representation ensembles, using neural networks to combine representations from language models and KG. This methodology allows for a deeper understanding of the context and relationships between entities mentioned in the news, thus improving classification accuracy. However, new challenges emerge from these approach, such as the risk of information loss in the graphs, due to the difficulty in leveraging the most relevant neighbors because of the large number of connections a node can have, but the opportunities of improve the impressive LLMs by significantly lower cost result in a wonderful opportunity. Rincon-Yanez et al.[19] propose enhancing KGs by generating inferred relationships and synthetic triples, improving representation and embedding accuracy, which can be adapted for fake news detection.

In the case of Yu-Hsianget al.[20] develops a model that employ KG to predict multiple tail or head entities for a given relation and entity, employing the relevant neighbors of the entities utilizing KG information in specialized or personalized domains. Meanwhile the work by Schmidt et. al. [21], integrates neural and symbolic methods to generate structured representations -KG- serving as external knowledge bases for RAG, in a similarly way, the course by Kolleger [22] provides a guide on how to use KGs in LLMs, where a RAG is used in conjunction with an advanced Cypher query language to improve information retrieval by LLMs by inputting external knowledge in the form of KG embeddings included in the input queries, which is a cheaper kind of fine-tuning since it does not require retraining the entire model.

Neuro-symbolic AI, as a combination of neural and symbolic methods, positions itself as a promising candidate for industrial applications. One benefit of neuro-symbolic solutions includes the integration of domain knowledge, e.g., in the form of Knowledge Graphs (KGs). Integrating KGs as a structured and symbolic knowledge representation into RAG-type applications offers a powerful approach to addressing the challenge of reducing hallucinations by combining the ability of language models to analyze text with the capability to retrieve relevant information from external sources, such as knowledge bases [21].

For instance, Ness's work introduces a framework for similarity search in massive graphs, which combines structure and content through the vectorization of nodes based on their neighbors' information. This method avoids costly isomorphism calculations (NP-complete problem useful for disambiguation), enabling efficient and scalable approximate searches in large networks [23].

C. Retrieval-Augmented Generation (RAG) for Enhanced NLP

Integrating KG into Retrieval-Augmented Generation (RAG) combines the ability of language models to analyze text with the capability to retrieve relevant information from external sources, such as external knowledge bases, to enhance accuracy and reliability, mitigating hallucinations [24]. RAG methods can be divided into two main categories: untrained and trained-based. The former directly exploit the knowledge retrieved during inference but do not fully optimize its usage, which can limit their effectiveness. In contrast, the trained-based methods adjust both the retriever and the generator, improving the integration of external information. Within this category, there are three main approaches: independent training, sequential training, and joint training[25] .

An example of joint training is the work by Yuhuai Wu and Markus [26], introduced a kNN-Augmented Attention Layer, where the model can retrieve relevant key-value pairs from external memory, integrating this information through local attention. Similarly, EAE [27] incorporates a specialized memory layer that allows the model to access specific representations for each entity mentioned in the text, learned directly from the text rather than relying on pre-trained external sources. Meanwhile, TOME [28] uses an external memory that stores dense representations of entity mentions extracted from Wikipedia and incorporates a MemoryAttention layer to integrate and retrieve specific entities. This approach enables the model to retrieve and reason about detailed, specific information from multiple textual sources without the need to reprocess the original text, facilitating both scalability and knowledge updating.

D. Use of Embeddings in Large Language Model

Embeddings are a type of data representation where words, phrases, or even entire entities are mapped to vectors in a continuous vector space. These vectors capture the semantic relationships between words and entities, enabling models to perform operations such as analogy reasoning, similarity matching, and knowledge retrieval[29]. Early embedding methods, such as Word2Vec and GloVe, primarily focused on word-level representations. However, more recent approaches have extended this idea to capture complex relationships between entities and their attributes in structured forms, such as KGs.

In the context of LLMs, embeddings play a crucial role by improving the models' ability to reason about relationships, contextualize inputs, and generate more accurate and relevant outputs. While embeddings such as BERT's token embeddings and GPT's attention mechanisms capture linguistic information, they often lack access to structured external knowledge, which is where Knowledge Graph Embeddings come into play. KAPALM : "Knowledge grAPh enhAnced Language Model for Fake News Detection"[30] and [31] "An efficient knowledge graph-based model for multi-domain fake news detection" highlight the importance of capturing semantic relationships and using external knowledge sources to enhance fake news detection. These works employ techniques such as Rotate, TransE, or DisMult to generate embeddings and enrich knowledge representation. They leverage pre-trained language models and the efficient construction of KG to provide a solid foundation in knowledge representation, thereby contributing to the accuracy and effectiveness of fake news detection.

IV. Preliminaries

A. Overview of Machine Learning for Fake News Detection

1) Datasets presentation and Selection: For this study, the WELFake [32] dataset was chosen, consisting of 72,134 news articles, with 35,028 real news and 37,106 fake news articles. The dataset contains four columns: Serial number (starting from 0), Title (the headline of the news article), Text (the content of the news article), and Label (a binary indicator, 0 for fake news and 1 for real news). This dataset was selected for its deterministic nature, where the classification task is straightforward with clear binary labels. Compared to other candidate datasets like HaluEval, which is more complex and focused on hallucination detection, or LIAR, which is centered on statement verification, WELFake provides a focused and scalable resource for detecting fake versus real news, making it particularly suitable for this study.

B. Knowledge Graph

A Knowledge Graph (KG) is a directed network model used to represent knowledge at various levels of abstraction and granularity. KGs consist of nodes or entities and edges, representing semantic relationships between entities. There are two main types of KGs: homogeneous, where all edges represent the same type of relationship, and heterogeneous, where edges can have different types of relationships or labels [33]. In both cases, a KG can be represented as a series of triples in the form (s, p, o), where s and o are entities and p are the relation connecting them. This model allows knowledge to be organized in a structured manner, facilitating human interpretation and automated analysis for tasks such as inference and semantic search [21].

C. Knowledge Graph Construction

The construction of knowledge graphs is key to organizing and structuring large volumes of information from unstructured data. One common approach to building these graphs is through the extraction of triplets of the type subject, predicate, object (S, P, O). This can be achieved using natural language processing (NLP) tools specialized in named entity recognition (NER) and relation extraction. Two popular approaches are Stanford CoreNLP, through the *Stanza* library, which allows these tasks to be performed sequentially, and REBEL (Relation Extraction By End-to-End Language Generation), which uses pre-trained language models like BART to simultaneously extract entities and relationships in a single step, thus improving efficiency and accuracy in the construction of the knowledge graph.

Additionally, recent methodologies leverage FAIR principles to enhance knowledge graph construction by ensuring findability, accessibility, interoperability, and reusability of data. Some approaches integrate NLP techniques with Open Knowledge Graphs (OpenKG), such as DBpedia or WikiData, to refine and enrich extracted triples, improving the accuracy and completeness of the generated graphs[34].

D. Knowledge Graph Embedding Techniques (TransE, RotatE, DistMult, ConvE)

An approximation to exploit the KG is by embedding his nodes into vectorial space; the techniques for these can be grouped into three main families based on their geometric approach and how they model relationships and entities according to [35]:

1) **Geometric Models**: The TransE model is one of the most well-known in this category. It uses a translation-based approach to model relationships. For a valid triplet (h, r, t), the embedding of the "tail" entity should be close to the embedding of the "head" plus a vector representing the relation r in the embedding space:

$$\mathbf{h} + \mathbf{r} \approx \mathbf{t}.$$

This model naturally captures hierarchical and structural relationships, making it suitable for many knowledge graphs. However, its simplicity presents limitations when modeling more complex relationships, such as non-transitive or many-to-many relations [36]. RotatE, also part of the **Geometric Models**, introduces a more sophisticated approach by modeling relationships as rotations in the complex plane. For the triplet (h, r, t), RotatE uses a rotation operation to transform the embedding of entity h into that of t:

$$\mathbf{t} \approx \mathbf{h} \circ \mathbf{r},$$

where $\circ$ denotes the Hadamard product. This strategy is particularly useful for modeling complex patterns such as symmetry, asymmetry, transitivity, and inversion [37].

2) **Convolutional Models**: ConvE is found in this family. This model uses convolutional neural networks to learn the embeddings, applying convolutions to the representations of the triplets to capture more complex patterns in the data. ConvE has proven particularly efficient at modeling knowledge graphs with relationships that require a deeper understanding of the interactions between entities. It uses an architecture based on a convolutional network that combines the representations of entities and relations, allowing for greater flexibility in capturing nonlinear interactions [25]. ConvE has successfully overcome the limitations of simpler models like TransE and DistMult by capturing more complex and nonlinear relationships in large knowledge graphs.

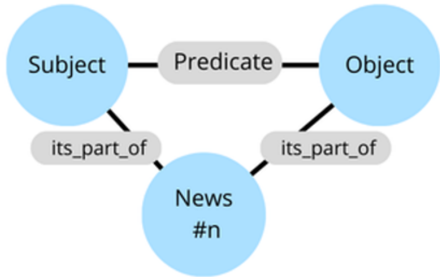

Fig. 1. Knowledge Graph model:Subject and Object nodes extracted using Stanford or REBEL methods, connected via a Predicate relationship.Each Subject-Predicate-Object (SPO) triplet is further linked to a Parent Node representing the news article from which the triplet was derived.

This classification highlights key differences based on empirical evidence. TransE is the most intuitive and computationally efficient model, using simple translations with fewer parameters. RotatE offers more flexibility by modeling complex relationships through rotations but is computationally more expensive.

E. Retrieval-Augmented Generation (RAG) Overview

The RAG model consists of two essential components: the Retriever, responsible for identifying relevant documents or text fragments from an external corpus using specific queries or similarity metrics based on embeddings [38]. The second component, the Augmented Generator, processes the retrieved information with the LLM to generate informed and contextualized responses, supporting its predictions with explicit evidence from external sources [39]. This reduces reliance on pre-trained parameters and enables real-time expansion of the knowledge base.

CommunityKG-RAG [40] incorporate knowledge graphs (KG) to improve retrieval and generation. By structuring KGs into communities using the Louvain algorithm, subsets of highly interconnected entities are identified, representing specific topics such as "presidential elections" or "climate change." These communities enhance precision in aligning claims with relevant topics through Sentence-BERT embeddings. Although incorporating all communities improves performance, the computational cost makes identifying relevant subsets crucial for efficiency, especially in multi-topic domains.

V. RAGS FOR FAKE NEWS DETECTION

The truthfulness of a news story will be validated based on the statements that compose it. To obtain

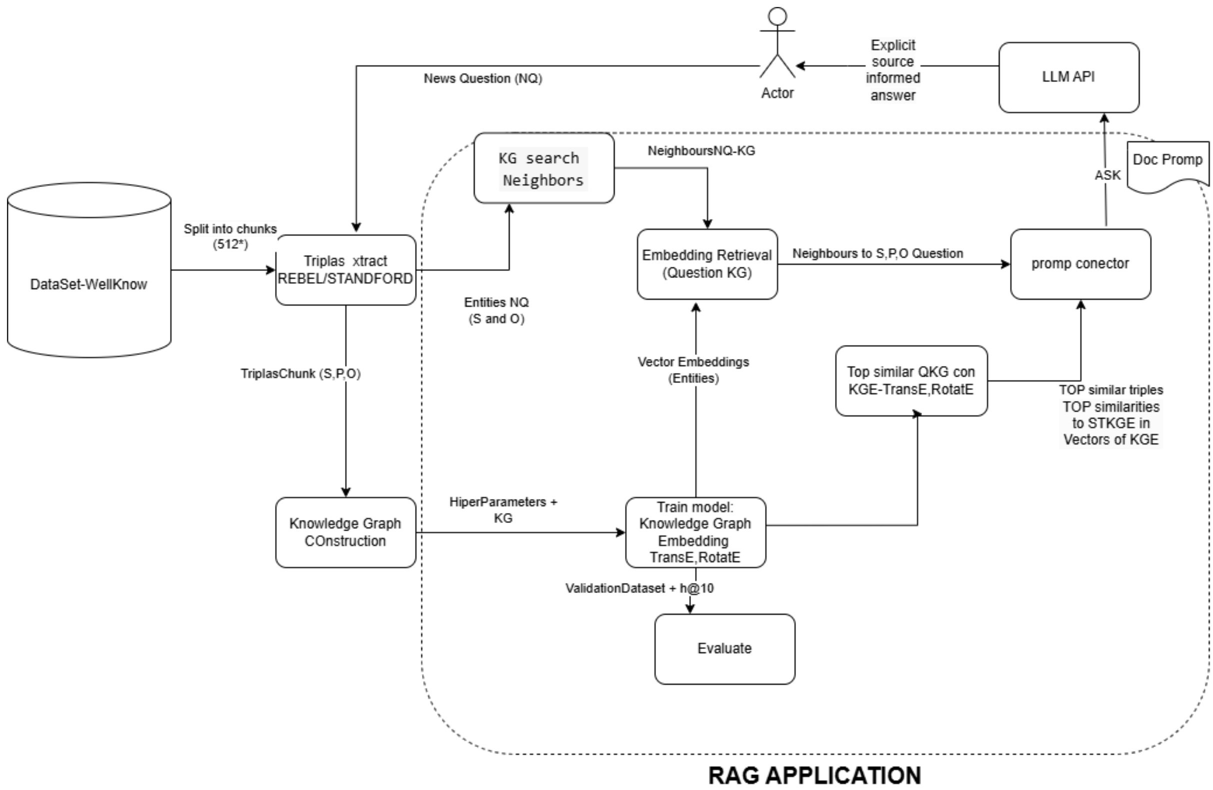

Fig. 2. Method description.

these statements, it is possible to use Named Entity Recognition (NER) and their linking, save them in a graph, and then compare them with reliable and truthful knowledge to verify their accuracy against a knowledge source that is sufficiently comprehensive for this purpose. Considering the tools currently available to humanity, LLMs stand out for their well-known capabilities in natural language processing, which are widely exploited and utilized. However, due to the inherent nature of the knowledge they acquire during training, they are limited to discussing what they learned during that phase, in addition to potentially generating false information and conclusions in the form of hallucinations. Therefore, to move closer to a scenario where these LLMs can be confidently used for fake news detection, they must be equipped with tools for this purpose. Specifically, using RAGs (Retrieval-Augmented Generation) with knowledge graph embeddings is explored, as they provide an ingenious option without requiring large financial, computational, or time costs.

A. *Knowledge Graph Construction*

The proposed method for constructing knowledge graphs (KG) involves advanced natural language processing (NLP) tools to extract entities and relationships from unstructured text data. To achieve this, Stanford CoreNLP, through the Stanza library, provides pre-trained models for tasks such as syntactic analysis, named entity recognition (NER), and relation extraction. Based on deep neural network architecture, Stanza specializes in extracting entities and relationships using a pipeline-based approach, where each task (NER, dependency parsing, relation extraction) is performed in sequential steps [41]. Stanza utilizes a pipeline-based approach to structure data into subject-predicate-object (S, P, O) triplets, forming the foundational elements of the knowledge graph. On the other hand, KG can also be generated using the REBEL model, as demonstrated in the work of Roos Bakker [42]. REBEL reformulates the task as a sequence-to-sequence (seq2seq) problem within a pre-trained language model, BART (Bidirectional and Auto-Regressive Transformer) [7]. Rather than performing separate tasks of entity recognition (NER) and relation classification, REBEL enables simultaneous extraction of both in a single step, significantly improving efficiency and accuracy in identifying entities and their relationships [16]. The main idea of how the graphs were generated can be observed in Fig 1.

Once the KG is constructed, various manipulations can be performed to extract information from it. A particularly simple but effective approach is the use of a neighborhood function, which will search for the nearest neighbors in proximity, with Sjikstra's algorithm for example. As highlighted by Arijit Khan (2023) [43], a detailed understanding of the relationships between entities within knowledge graphs improves the ability to answer questions more accurately, particularly when questions involve complex concepts or require multi-step reasoning. By incorporating such neighborhood information, we can bridge the gap between surface-level entity recognition and deeper, more accurate understanding of news.

B. Knowledge Graph Embedding Techniques (TransE, RotatE, DistMult, ConvE)

The graphs generated using Stanford NLP and REBEL were processed to obtain embeddings in the vector space. Two Knowledge Graph Embedding (KGE) techniques were then applied to capture the semantic relationships and structure of the graphs:

- **TransE:** A translational model that represents relationships as translations in vector space, ideal for hierarchical and transitive relationships.
- **RotatE:** An extension of TransE that models relationships through rotations in complex space, enabling the capture of symmetries and asymmetries.

C. Nearest entities and Relation Prediction

For the search of nearest entities, the unique entities extracted from the subject and object nodes of the news has been mapped into the vector space of the global generated KGE wich is the knowledge base. Cosine similarity was computed between the vector of the target entity and the vectors of all other entities in the vector space. This step allowed for the identification of entities most similar to the target entity in the news. Subsequently, for each combination of the subject entity from the news and its nearest neighbors, the KGE model was used to predict the most probable relation between them. This step enabled the identification of potential connections between the news and the existing knowledge base.

D. RAG - Augmented Knowledge Generation:

Finally, the extracted entities, relationships, and triplets were used as augmented knowledge. This enriched context was then fed into the language model (LLM) during the final stage to generate predictions or grounded responses based on the provided knowledge.

VI. Experiment

In this section, we describe the methodology used to evaluate our Retrieval-Augmented Generation (RAG) model on the task of fake news detection. The model propose integrates knowledge graph (KG) embeddings to enhance the retrieval process, aiming to improve the accuracy and reliability of predictions regarding the truthfulness of news articles. The evaluation is designed to assess both the performance of the RAG model and the quality of the generated text regarding factual accuracy, contextual understanding, and transparency.

A. Evaluation Setup

The *setup* used was hybrid[44], combining cloud processing and local resources:

- **Cloud:** Google Colab with TPU was used for processing the REBEL model, generating RotatE and TransE *embeddings*, and leveraging APIs for large language models (LLMs).
- **Local-earth:** Stanford CoreNLP (Stanza) was executed on a GPU, NVIDIA GeForce RTX 3050.

B. Experimental Design

The experimental design consists of five main stages, each focused on a critical aspect of the fake news validation pipeline using a Retrieval-Augmented Generation (RAG) model with Knowledge Graph Embeddings (KGE), the whole process can be view in the Fig 2. and its describe as these:

1) **Extraction of Triplets from the News Article:** The news article to be validated was processed following the same approach used in constructing the knowledge base. This involved extracting triplets in the `(subject, predicate, object)` format using the previously described techniques. This step ensured that the semantic structure of the news article was properly integrated into the pipeline, facilitating the subsequent search for related entities. An example of this setup can be seen in Figure 2

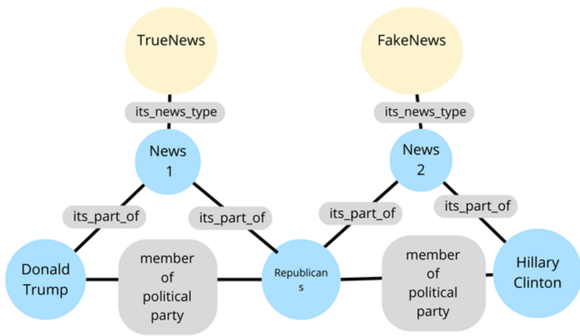

Fig. 3. Example of a graph generated from two news articles: Entities are connected to their respective news articles and their parent nodes.

2) **Extraction of Closest Neighbors Directly Related from the Knowledge Graph:** In this step, the closest neighbors directly related to the `subject` and `object` entities were identified within the knowledge graph. This was done using the shortest path algorithm (Dijkstra), applied to find the most relevant neighbors in the graph based on direct relationships, as result triples dataset have been obtained.
To generate the nearest entities and their relations in the *Knowledge Graph* (KG), the **NetworkX** library was employed. A directed graph representing the triplets `(head,`

Scenarios Configuration	Extraction	Neighbours Search	Embeddings	KBContext
Knowledge Base (KB) + Shortest Path Length (SPF)				
Standford	X	-	-	Neighbours
Rebel	X	-	-	Neighbours
KB + KGE				
Standford	-	-	TransE	Predicted
Standford	-	-	RotatE	Predicted
Rebel	-	-	TransE	Predicted
Rebel	-	-	RotatE	Predicted
KB + SPF + KGE				
Standford	X	-	TransE	Predicted + Neighbours
Standford	X	-	RotatE	Predicted + Neighbours
Rebel	X	-	TransE	Predicted + Neighbours
Rebel	X	-	RotatE	Predicted + Neighbours
Base Question Direct				
None	-	-	None	-

TABLE I
SCENARIOS CONFIGURATION

`relation, tail)` was initially constructed. Subsequently, the 10 closest neighbors with direct relations were identified using the `single_source_shortest_path_length` function (which implements Dijkstra's algorithm), starting from the entities corresponding to the query news articles.

3) **Search for the Closest Neighbors in the Knowledge Graph Embedding (KGE):** Two embedding models, TransE and RotatE, were trained on a selected news dataset, which was split into training, testing, and validation sets. The models were evaluated based on metrics such as hits and mean reciprocal rank (MRR), and the best-performing configurations were used to obtain the Knowledge Graph Embedding (KGE) for each constructed graph, specifically for the Stanford and REBEL graphs. Using the unique entities extracted from the `subject` and `object` nodes of the news articles, a search was conducted in the vector space of the KGE. This procedure involved:

   a) **Loading Embeddings:** The embeddings for the entities were loaded from the trained KGE model using entity and relation identifiers (`entity_to_id` and `relation_to_id`), creating a dictionary of embeddings where each entity was mapped to its corresponding vector representation.
   b) **Cosine Similarity Calculation:** Cosine similarity was used to measure the proximity between the target entity vector and the vectors of all other entities in the graph.
   c) **Top-k Neighbor Selection:** The similarity values were ordered to identify the most similar entities, excluding the target entity itself. The *top k* closest neighbors were selected using the `nearest_neighbors` function, implemented as follows:
      * The cosine similarity between the embedding of the target entity and the embeddings of all other entities was computed.
      * The results were sorted in descending order of similarity.
      * The top *k* closest entities were returned.

4) **Prediction of Relationships:** For each combination of the entities of the news article and its closest neighbors, the KGE model was used to predict the most likely relationship between them. This step enabled the identification of potential connections between the news article and the existing knowledge base, adding contextual validation to the predictions generated.
5) **Generation with Augmented Knowledge:** The entities, relationships, and contexts obtained in the previous steps were combined to form an enriched set of data. This augmented knowledge was sent to the LLM ,Api Chatgpt, in the final stage, enabling the generation of predictions and responses based on explicit and contextual evidence.

C. Evaluation

The experimental evaluation is based on 11 distinct scenarios designed to test the RAG model's performance under various conditions. These scenarios are categorized by the method used for entity and relationship extraction (REBEL or Stanford) and the information retrieved and sent to the LLM. In all cases, the knowledge base (KB)—the news article being validated—is included as input. The scenarios then vary based on specific criteria: whether the neighbors of the entities in the news article are sent (SPF), the type of embedding used to retrieve the closest entities

(KGE), and the combination of all elements. The most comprehensive scenario involves sending SPF + KGE + KB. These scenarios are detailed in Table I.

.The best model configurations were RebelTransE1, StandfordTransE1, RebelRotatE1, and StandfordRotatE1. Notably, training with the Stanford knowledge graph took six times longer than with Rebel, using the same configuration, without guaranteeing better results.Table Table II presents the obtained values for the Mean Inverse Rank (MIR) and Hits@K metrics, comparing different model configurations

a) Metrics: To evaluate the performance of the model across these scenarios the following metrics was considered:

1) **Precision:** This metric evaluates the model's ability to identify true or false news articles correctly. Precision is computed as the ratio of true positive predictions (correct classifications) to predicted positive instances. A high precision score indicates that the model makes reliable predictions about the veracity of news articles, minimizing false positives.
2) **Accuracy:** It is a metric that validates the number of correct statements (true or false) obtained in the response based on the triplets. In other words, the metric evaluates how many parts of the response identified as true or false are correct, compared to the real facts represented in the triplets extracted from the dataset.
3) **Contextual Understanding (Hallucinations):** This metric measures the model's ability to avoid generating irrelevant or nonsensical responses (hallucinations). It evaluates how well the model maintains contextual relevance when generating responses. Hallucinations often occur when the model fails to use external knowledge effectively or generates fabricated information.
4) **Compelling Misinformation:** This metric assesses the amount of incorrect but highly convincing information generated by the model. Misinformation can undermine the credibility of a fake news detection system, and this metric helps evaluate the model's ability to avoid generating content that may seem plausible but is factually incorrect.
5) **Transparency and Traceability:** Transparency evaluates the model's ability to cite relevant sources and provide a clear path for tracking the origin of the information used in the decision-making process. In a fake news detection context, this is crucial to ensure that the model's conclusions can be verified by external users, particularly in academic or journalistic settings, employing the Likert scale.
6) **Source Retrieval Accuracy:** Given that our model relies on the retrieval of external knowledge sources (documents and KG embeddings), this metric evaluates how accurately the model retrieves relevant information. A higher retrieval accuracy ensures that the model has access to the most pertinent data, which can lead to more informed and accurate news verifications.

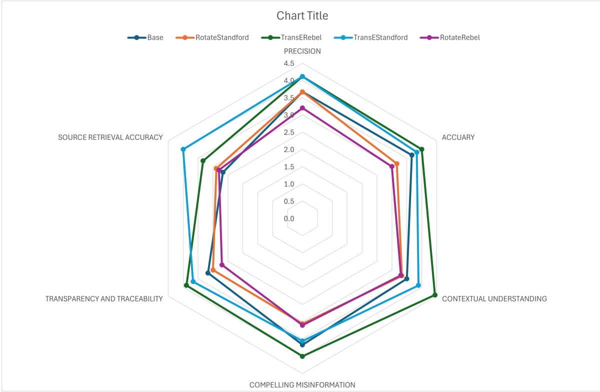

Fig. 4. Experimental results.

VII. Results

A. Comparison of Results

a) Stanford NLP: The analysis conducted with Stanford NLP yielded a total of **560,630 triples**, derived from **163,008 unique entities** and **46,559 unique relations**. Although this model identifies a large number of triples, many of the most frequent entities are personal pronouns, such as *he*, *it*, or *they*, which might suggest the need for further filtering to reduce noise in the data.

b) REBEL: In comparison, the REBEL model produced **54,358 triples**, with **25,250 unique entities** and **228 unique relations**.

Despite generating fewer data points than Stanford NLP, the REBEL results seem more concise and relevant to the analyzed domain. For example, the most frequent entities include political figures and key terms like *Donald Trump*, *Hillary Clinton*, and *Republican*, while in Stanford, the most common relation is *position held*, reflecting a more semantic approach and less reliance on explicit hierarchical structures as in Stanford NLP.

B. Baseline Model vs Proposed RAG Approach

Six metrics were applied to evaluate the tested models' performance, as described in Section 5. experiment: precision, accuracy, contextual understanding, compelling misinformation, transparency and traceability, and source retrieval accuracy.

Ten experiments were conducted for each case to validate the truthfulness of a news item unknown to the embedding model, using five true and five false news items.

The average of the evaluations of the scenarios in each metric was used to rate the proposed approach, obtaining as a result Fig. 5.Figure 4. The results

TABLE II
EXPERIMENT RESULTS WITH TRANSE AND ROTATE MODELS

Experiment	Loss	Dimention	Lear. Rate	Epochs	Hits@1	Hits@3	Hits@5	Hits@10	MRR
TRANSE									
RebelTransE1	NSSA	100	0.002	200	0.0593	0.3751	0.4384	0.5034	0.2369
RebelTransE2	Softplus	70	0.001	50	0.0562	0.3108	0.3656	0.4184	0.1992
RebelTransE3	Softplus	30	0.001	50	0.0536	0.2959	0.3427	0.3873	0.1867
RebelTransE4	Softplus	100	0.0001	25	0.0099	0.2517	0.2833	0.3129	0.1391
StandfordTransE1	Softplus	70	0.001	50	0.305	0.4253	0.4622	0.4983	0.3748
StandfordTransE2	Softplus	150	100	250	0.0682	0.0944	0.1064	0.1226	0.0879
StandfordTransE3	Softplus	100	0.001	150	0.0562	0.3108	0.3656	0.4184	0.1992
StandfordTransE4	Softplus	30	0.0001	25	0.0325	0.0647	0.0786	0.0741	0.0628
ROTATE									
RebelRotatE1	marginranking	300	0.001	100	0.1406	0.1918	0.2167	0.2534	0.1796
RebelRotatE2	Softplus	50	0.05	100	0.0682	0.0943	0.1064	0.1226	0.0879
RebelRotatE3	marginranking	70	0.01	100	0.0857	0.1158	0.1318	0.1552	0.1095
RebelRotatE4	Softplus	150	0.0001	100	0.0101	0.0204	0.0315	0.0522	0.0247
StandfordRotatE1	marginranking	300	0.001	100	0.2812	0.306	0.3173	0.3325	0.3006
StandfordRotatE2	marginranking	500	0.001	500	0.1959	0.272	0.2979	0.3371	0.2452
StandfordRotatE3	Softplus	800	0.001	900	0.1277	0.182	0.2084	0.2407	0.1664
StandfordRotatE4	Softplus	450	0.001	400	0.1257	0.184	0.2088	0.2423	0.1654

TABLE III
EXPERIMENT RESULTS

Metrics	Base	Standford		REBEL	
		RotatE	TransE	RotatE	TransE
PRECISION	3.7	3.7	4.1	3.2	4.1
ACCURACY	3.7	3.2	4.0	3.0	3.8
CONTEXTUAL UNDERSTANDING	3.5	3.3	4.4	3.3	3.9
COMPELLING MISINFORMATION	3.7	3.1	4.0	3.1	3.6
TRANSPARENCY AND TRACEABILITY	3.2	3.0	3.9	2.7	3.7
SOURCE RETRIEVAL ACCURACY	2.7	2.9	3.3	2.8	4.0
Total	3.4	3.2	4.0	3.8	3.0

can be seen in Table III, where it is evident that the best-performing model is TransE with REBEL. However, in terms of precision—specifically, accurately predicting whether a news item is false—it ties in with TransE and Stanford, which also achieved the best performance in retrieving useful information for decision-making. On the other hand, the worst-performing model overall was RotatE with REBEL, although it predicted truthfulness just as well as the baseline model.

The most computationally and time-intensive model was RotatE with Stanford, requiring more than 8 hours of training on a Google Colab T4 GPU. In all cases, the models tended to predict true news correctly but often misclassified false news.

C. *Analyzing the results: An example from the dataset*

A case where almost all models made an error was one involving WikiLeaks, a news item supposedly generated by the Washington Examiner. This led the baseline model to trust the credibility of this outlet and assume the news was true. It responded, "Additionally, the reference to the source 'WE' indicates that the information may be traced back to a reputable news outlet." In contrast, the complete RAGs, embeddings, and KG neighbors prioritized the entities mentioned in the case, such as Clinton, Podesta, and Doug Band, since they were found in both retrievals. However, the model that directly sent the neighboring entities from the graph made the correct prediction by relying on the fact that the relationships between the entities in the news and their neighbors did not suggest the behavior indicated by the news.

VIII. CONCLUSIONS AND FUTURE WORK

A. *Summary of Findings*

While the LLMs can model the veracity of news, they rely on the assumption that if the supposed source is a real organization, such as a magazine, government organization, or a recognized person, everything they say must be true. This leads to hallucinations when concluding these false facts. For example, in a news story about WikiLeaks, the model erroneously claims the news is true because it mentions Clinton, Podesta, and the source Washington Examiner (WE). Still, it doesn't consider the content of the news or what it knows about these entities. However, this also happens with the complete model, which makes the same error by classifying it as true, as it has no sources contradicting it, and the entities in the news are close.

In the experiment that uses the nearest neighbors with their relationships, extracted solely from the knowledge graph, the model correctly predicts the news as false, inferring that since there are no specific relationships between the entities that describe behavior like the one mentioned in the news, the news must be false. In this case, for the complete experiment, the excess of information generated noise in the response, prioritizing inference based on proximity. The quality of the retrieval is directly affected by the embedding model's training quality.

B. Limitations of the Current Approach

News with high semantic content but few entities (people, places, or things) tend to perform worse, sometimes causing hallucinations due to lack of context, particularly in REBEL. This issue is significantly reduced with Stanford due to a higher granularity in the extracted triplets. However, this comes at the cost of generating embeddings, as it quadruples the number of entities obtained compared to REBEL.

C. Future Directions for Research

Future work will focus on validating the proposed model using evaluation metrics such as F1-score, accuracy, and area under the ROC curve (AUC-ROC). These metrics will provide a comprehensive assessment of the model's performance and its ability to accurately differentiate between classes. Additionally, hyperparameter tuning and the inclusion of additional data will be explored to further optimize the model's effectiveness.

On the one hand, further exploration is needed in manipulating knowledge graphs to obtain richer knowledge and establish more complex relationships between entities. On the other hand, more families and methods for generating embeddings may yield better results, which would also allow for expanding the proposed RAG methodology to other scenarios.

I also believe this information retrieval method can be exploited in scenarios like zero-shot questions or chatbots.

As future work, the use of federated approaches will be explored to optimize the execution of queries on large knowledge graphs. In particular, its application in Retrieval-Augmented Generation (RAG) models for fake news detection would facilitate the processing of large data volumes and the extraction of complex relationships in real-time. Additionally, the impact of this approach on reducing operational costs and the sustainability

Furthermore, I see great potential in embedding information into vector spaces for LLMs regarding information security, as it prevents the exposure of easily exploitable data, unlike when information is in flat form. This could address one of the current problems companies face regarding the potential exposure of sensitive information through cyberattacks.

REFERENCES

[1] G. M. Nieves-Cuervo, E. F. Manrique-Hernández, A. F. Robledo-Colonia, and E. K. A. Grillo, "Infodemia: noticias falsas y tendencias de mortalidad por covid-19 en seis países de américa latina," *Revista Panamericana de Salud Pública*, vol. 45, p. 1, 5 2021.

[2] A. Alvarez-Risco, C. R. Mejia, J. Delgado-Zegarra, S. Del-Aguila-Arcentales, A. A. Arce-Esquivel, M. J. Valladares-Garrido, M. R. del Portal, L. F. Villegas, W. H. Curioso, M. C. Sekar, and J. A. Yáñez, "The peru approach against the covid-19 infodemic: Insights and strategies," *The American Journal of Tropical Medicine and Hygiene*, vol. 103, pp. 583–586, 8 2020.

[3] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, pp. 211–236, 5 2017.

[4] N. Dziri, S. Milton, M. Yu, O. Zaiane, and S. Reddy, "On the origin of hallucinations in conversational models: Is it the datasets or the models?," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, eds.), (Seattle, United States), pp. 5271–5285, Association for Computational Linguistics, July 2022.

[5] K. Nagy and J. Kapusta, "Improving fake news classification using dependency grammar," *PLOS ONE*, vol. 16, p. e0256940, 9 2021.

[6] G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu, "Can knowledge graphs reduce hallucinations in llms? : A survey," 2024.

[7] L. Berglund, M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans, "The reversal curse: Llms trained on "a is b" fail to learn "b is a"," 2024.

[8] J. Baek, A. F. Aji, and A. Saffari, "Knowledge-augmented language model prompting for zero-shot knowledge graph question answering," 2023.

[9] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–20, 2024.

[10] G. Di Paolo, D. Rincon-Yanez, and S. Senatore, "A Quick Prototype for Assessing OpenIE Knowledge Graph-Based Question-Answering Systems," *Information (Switzerland)*, vol. 14, p. 186, mar 2023.

[11] Y. Wang, L. Wang, Y. Yang, and Y. Zhang, "Detecting fake news by enhanced text representation with multi-edu-structure awareness," 2022.

[12] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, (New York, NY, USA), p. 395–405, Association for Computing Machinery, 2019.

[13] K. Shu, S. Wang, and H. Liu, "Understanding user profiles on social media for fake news detection," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 430–435, 2018.

[14] A. Jaradeh and M.-B. Kurdy, "Aremotive bridging the gap: Automatic ontology augmentation using zero-shot classification for fine-grained sentiment analysis of arabic text," *IEEE Access*, vol. 11, pp. 81318–81330, 2023.

[15] M. Krötzsch and V. Thost, "Ontologies for knowledge graphs: Breaking the rules," in *The Semantic Web – ISWC 2016* (P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, and Y. Gil, eds.), (Cham), pp. 376–392, Springer International Publishing, 2016.

[16] H. Li, "Make meaningful knowledge graph from opensource rebel model," *Medium*, 2023.

[17] Y. Dun, K. Tu, C. Chen, C. Hou, and X. Yuan, "Kan: Knowledge-aware attention network for fake news detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 81–89, 5 2021.

[18] B. Koloski, T. S. Perdih, M. Robnik-Šikonja, S. Pollak, and B. Škrlj, "Knowledge graph informed fake news classification via heterogeneous representation ensembles," *Neurocomputing*, vol. 496, pp. 208–226, 7 2022.

[19] D. Rincon-Yanez, A. Mouakher, and S. Senatore, "Enhancing downstream tasks in Knowledge Graphs Embeddings: A Complement Graph-based Approach Applied to Bilateral Trade," *Procedia Computer Science*, vol. 225, pp. 3692–3700, 2023.

[20] Y.-H. Lin, H.-T. Shieh, C.-Y. Liu, K.-T. Lee, H.-C. Chang, J.-L. Yang, and Y.-S. Lin, "Retrieval-augmented language model for extreme multi-label knowledge graph link prediction," 2024.

[21] W. Schmidt, D. Rincon-Yanez, E. Kharlamov, and A. Paschke, "Scaling Scientific Knowledge Discovery with Neuro-Symbolic AI and Large Language Models," in *Joint Proceedings of Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems co-located with 20th International Conference on Semantic Systems (SEMANTiCS 2024)*, vol. 3759 of CEUR Workshop Proceedings, 2024.

[22] A. Kolleger, "Knowledge graphs for rag."

[23] A. Khan, N. Li, X. Yan, Z. Guan, S. Chakraborty, and S. Tao, "Neighborhood based fast graph search in large networks," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, (New York, NY, USA), p. 901–912, Association for Computing Machinery, 2011.

[24] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021.

[25] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A survey on rag meeting llms: Towards retrieval-augmented large language models," 2024.

[26] Y. Wu, M. N. Rabe, D. Hutchins, and C. Szegedy, "Memorizing transformers," 2022.

[27] T. Févry, L. B. Soares, N. FitzGerald, E. Choi, and T. Kwiatkowski, "Entities as experts: Sparse memory access with entity supervision," 2020.

[28] M. de Jong, Y. Zemlyanskiy, N. FitzGerald, F. Sha, and W. Cohen, "Mention memory: incorporating textual knowledge into transformers through entity mention attention," 2022.

[29] D. Rincon-Yanez, C. Ounoughi, B. Sellami, T. Kalvet, M. Tiits, S. Senatore, and S. B. Yahia, "Accurate prediction of international trade flows: Leveraging knowledge graphs and their embeddings," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, p. 101789, dec 2023.

[30] J. Ma, C. Chen, C. Hou, and X. Yuan, "Kapalm: Knowledge graph enhanced language models for fake news detection," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3999–4009, 2023.

[31] L. fu, H. Peng, and S. Liu, "Kg-mfend: an efficient knowledge graph-based model for multi-domain fake news detection," *The Journal of Supercomputing*, vol. 79, pp. 18417–18444, 11 2023.

[32] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "Welfake: Word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 8, pp. 881–893, 8 2021.

[33] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," *ACM Computing Surveys*, vol. 54, pp. 1–37, 5 2022.

[34] D. Rincon-Yanez and S. Senatore, "FAIR Knowledge Graph construction from text, an approach applied to fictional novels," *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with 19th Extended Semantic Conference (ESWC 2022)*, vol. 3184 of CEUR Workshop Proceedings, pp. 94–108, 2022.

[35] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering," *ACM Computing Surveys*, vol. 31, pp. 264–323, 9 1999.

[36] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, Curran Associates, Inc., 2013.

[37] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," 2019.

[38] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021.

[39] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," 2020.

[40] R.-C. Chang and J. Zhang, "Communitykg-rag: Leveraging community structures in knowledge graphs for advanced retrieval-augmented generation in fact-checking," 2024.

[41] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," 2020.

[42] D. D. S. a. Roos Bakker a, b and M. de Boer, "Knowledge graph extraction in a practical context," in *Neurosymbolic Artificial Intelligence*, 2022.

[43] A. Khan, "Knowledge graphs querying," 2023.

[44] D. Rincon, C. Diaz, O. Garcia, and H. Bolivar, "Federated cloud architecture model using opportunistic university resources," *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, no. E17, pp. 198–211, 2019.