2024-09-17

An Exploration of Causality in Social Media Data with Knowledge Graphs

Ravi, Rahul

UNIVERSITY OF CALGARY

An Exploration of Causality in Social Media Data with Knowledge Graphs

by

Rahul Ravi

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN COMPUTER SCIENCE

CALGARY, ALBERTA

SEPTEMBER, 2024

ABSTRACT

This study explores the integration of Knowledge Graphs (KGs) with Large Language Models (LLMs) to perform causal analysis on text-based social media data. The objective is to uncover the underlying causes and sentiments driving discussions around emergency management scenarios such as the Israel-Palestine conflict, thereby providing critical insights for decision-making. The research focuses on advanced techniques to effectively represent text as KGs and retrieve the most suitable context of information to analyse. Various methods are evaluated across different datasets. The proposed model, PRAGyan, combines LLMs and KGs under the Retrieval Augmented Generation (RAG) framework. It utilizes the Neo4J Graph Database to handle continuous real time data and GPT-3.5 Turbo LLM for causal reasoning. This yields more accurate results compared to the baseline model (GPT-3.5 Turbo LLM without KG). Quantitative analysis using metrics such as BLEU and cosine similarity show an improvement by 10%.

PUBLICATION

Rahul Ravi, Gouri Ginde, and Jon Rokne, "PRAGyan - Connecting the Dots in Tweets", *Accepted at the 16th International Conference on Advances in Social Networks Analysis and Mining -ASONAM-2024.*

Summary: The integration of Knowledge Graphs (KGs) and Large Language Models (LLMs) for performing causal analysis on social media data, particularly tweets has been explored. The study aims to understand the reasons behind events and statements in social media by leveraging the strengths of KGs for their structured relational knowledge and LLMs for their deep contextual understanding. A hybrid model is proposed, PRAGyan, which uses Retrieval-Augmented Generation (RAG) to enhance the causal reasoning capabilities of LLMs by incorporating relevant context from KGs. The methodology involves extracting causal information from tweets about COVID-19 and evaluating the effectiveness of the proposed model against a baseline LLM (GPT-3.5 Turbo). The results indicate that the PRAGyan model outperforms the baseline in both qualitative and quantitative metrics, demonstrating the benefits of integrating KGs with LLMs for improved interpretability and actionable insights in social media analysis.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude towards my late supervisor Dr. Guenther Ruhe. His guidance and encouragement helped me navigate through the difficult initial stages of my program and helped me devise a compelling research problem to focus my efforts on. I miss him dearly. I hope the work that followed would make him proud.

I am also extremely thankful to my supervisors, Dr. Jon Rokne and Dr. Gouri Ginde, without whose support and mentorship this research would not have been possible. Their critical feedback through out the past year helped me develop sharper thinking and hone my technical skills. Dr. Rokne's years of experience and attention to detail made me a better academic writer. I am truly honoured to have been his student. Dr. Ginde's passion for research and vast knowledge in the field compelled me to work harder and realise my true potential. She volunteered to mentor me when I struggled to find a new supervisor and has always been looking out for me ever since.

I am very grateful to Dr. Tyler Bonnell and Dr. Ahmad Abdel Latif for agreeing to serve on my defense committee and giving me the opportunity to experience such a significant event in my life, whose constructive criticism and feedback will guide me as I continue to improve and grow in the years to come.

And I would like to acknowledge the invaluable support provided by my parents and my brother over the years. Their unwavering presence, no matter the time of day, and the sacrifices they made to give me a good life have brought me to this point. I am deeply grateful for everything they have done.

Finally, I must recognize my best friend, Alisha. Her strength in overcoming life's challenges has been the driving force behind my pursuit of success in recent years. Her resilience has inspired me to keep pushing forward, no matter the obstacles. Without her by my side, I would not be the person I am today.

Contents

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Term	Definition
Knowledge Graph (KG)	A graph-based data structure representing relationships (as edges) between entities (as nodes).
Entities	Specific pieces of information or objects within texts that carry significant meaning and belong to predefined categories.
Named Entity Recognition (NER)	Identifying and classifying entities in text, such as names, organizations, and locations.
Natural Language Processing (NLP)	A field of AI focused on enabling computers to understand and process human language.
Large Language Models (LLMs)	Computational models trained on large text data that are capable of general-purpose language generation and various natural language processing tasks like classification.
Retrieval-Augmented Generation (RAG)	Combining retrieval mechanisms for obtaining contexts from text, with generative models.
Context	Relevant information pertaining to an event that can be extracted from a vast amount of data.
Graph Neural Networks (GNNs)	Neural networks designed to operate on graph-structured data, capturing semantic and relational information.
Contextual Similarity	Measures how closely two texts are related in meaning, regardless of differences in structure, wording, or length.
Cosine Similarity	A metric used to measure the similarity between two nonzero vectors.
Entity Embeddings	Representing entities as vectors to capture semantic and emotional contexts.
Temporal Dynamics	Changes in the relevance and context of social media posts over time.
Sentiment Analysis	Determining the sentiment expressed in text to understand emotions and opinions.
Node2Vec	A graph embedding technique using random walks to generate node embeddings.
Sentence-BERT	A modification of BERT fine-tuned for producing meaningful sentence embeddings.
BLEU (Bilingual Evaluation Understudy)	A metric for evaluating machine-translated text by comparing n-grams of the candidate and reference translations.
Dynamic Routing Graph Inference (DRGI)	A method for dynamic graph construction and updating based on routing algorithms.
Actionable Insights	Inferences or insights gather from text through reasoning based on a query or event that can lead to decision makers taking appropriate actions, in the context of emergency management scenarios.

Chapter 1

Introduction

Text-based data captured from various online sources has become ubiquitous in the modern world, primarily due to the advent of the internet and the subsequent rise of Big Data. Much of this data often goes unnoticed and unused, despite containing crucial information that could significantly enhance decision-making processes through causal analysis during crisis management and emergency response scenarios. Social networks, enabled by the internet, have become real-time information-sharing platforms from which large amounts of textual data can be collected. When irrelevant data is filtered out, this data may contain valuable information. One area where social media can be particularly helpful is in emergency scenarios. In such situations, real-time causal analysis through continuous monitoring and examination of events as they unfold is crucial. This allows for immediate awareness and reaction to changing conditions, yet it is very difficult to manage manually by domain experts due to the sheer volume and speed of data generation. Automating this process can lead to better management of time and resources, thereby mitigating potential threats more effectively [2]. For example, automated systems can quickly analyze social media posts to identify causal relationships, such as determining that a spike in emergency calls is due to infrastructure damage following an earthquake. This enables faster deployment of rescue teams to the most critical locations, reducing the time to reach and assist those in need. This research focuses on this aspect to uncover efficient techniques to understand the causes behind various phenomena observed in social media data and gain insights into the kind of sentiments expressed about them.

Natural Language Processing (NLP) is a widely used AI technique for processing text-based data and has been extensively researched in recent years. Innovations such as pre-trained models and transformers like BERT have spurred the development of new

technologies and tools, including Chat GPT [3]. One of the most common applications of NLP is sentiment analysis, which has a wide range of uses from e-commerce to brand marketing and social media analysis. Sentiment analysis helps in understanding the emotions behind a body of text, which can be invaluable in gauging public opinion, customer satisfaction, and more [4].

However, representing text in a structured manner for better processing and analysis can be challenging due to the unstructured nature of text data, which includes variability in language, context and ambiguity, and the complexity of capturing relationships between entities. Knowledge Graphs (KGs) and Ontologies offer a possible solution to this problem by providing structured representations of information and defining relationships between entities. KGs, in particular, are relatively new but they provide an efficient multi-graphical representation for easy querying and searching for relationships, patterns, and insights within large and complex datasets. They are considered one of the best models for databases due to their speed, low complexity, and comprehensive view of the stored data [5]. KGs can represent complex relationships between data points such as hierarchical, associative, temporal, causal, and contextual relationships, making them ideal for extracting meaningful insights from vast amounts of unstructured text data [6].

This research tries to leverage real-time social media text-based data, so that actionable insights, those that can directly inform and drive decision-making and prompt specific actions, can be extracted through the incorporation of KGs. The goal is to enhance the understanding and analysis of sentiments expressed about various factors pertaining to specific events in social media, which can be crucial for various applications, including crisis management, marketing strategies, and public opinion monitoring [7]. For instance, during a pandemic such as COVID-19, by leveraging real-time sentiment analysis of social media data, health authorities can monitor discussions about symptoms and outbreaks, allowing them to quickly identify emerging hotspots and allocate medical resources more efficiently to areas experiencing spikes in cases. Additionally, companies producing health-related products, such as masks and sanitizers, can analyze social media sentiment to gauge public demand and adjust their marketing strategies accordingly.

If there is a surge in positive sentiment towards certain types of masks, companies can increase production and focus their advertising on those products. Furthermore, governments can track public sentiment regarding lockdown measures and vaccination programs. By understanding public concerns and misconceptions, they can tailor their communication strategies to address specific issues, thereby increasing compliance and trust in public health directives.

Hence, it has been proposed in this research to perform and automate the process of causal reasoning. By identifying the right set of factors driving opinions and sentiments behind any given statement about an ongoing crisis scenario, it becomes easier for decision makers to take actions. Through the use of KGs, it is explored whether this can be achieved.

In this research, tweets were chosen as the primary data source, with open-source data obtained from the Kaggle platform. Tweets often contain hashtags that help categorize them by topic, allowing for effective filtering based on relevant subjects. When analyzed within a specific timeframe, these tweets can provide insights into public opinions on various issues, particularly within the same geographic region. By gathering and examining a collection of tweets surrounding a particular event or opinion, we can uncover patterns, correlations, and potential underlying causes of the event, offering valuable context and reasoning. The challenge lies in identifying the right set of tweets that relate directly to the primary tweet of interest. The objective of this research is to explore whether representing these tweet collections as knowledge graphs can enhance the extraction of such relevant information, making it easier to reason about the relationships and causes behind specific events or opinions.

The approach was began by representing an entire tweet as a KG node, with edges based on cosine similarity to other tweets. It incorporated sentiment analysis to facilitate root cause analysis of sentiments expressed in the tweets. By examining the connections between tweets, the aim was to identify patterns and commonalities that could explain the underlying sentiments [8], behind statements of interest to explore causality. Here, it was primarily manual observations of visualizations of the KG and metrics such as

centrality measures to infer causality.

This representation evolved to a more refined approach using Named Entity Recognition (NER) where entities within tweets were identified and represented as nodes in the KG, while depicting the relationships between these entities as edges [9, 10]. Sentiments were assigned as edge weights, providing a detailed view of how different entities were connected and how sentiments flowed between them [11]. Additionally, contextually similar entities were linked to enrich the KG. This method aimed to identify tweets that were most likely to explain causes behind the sentiment expressed in a given input statement, thus facilitating deeper insights into sentiment dynamics. A key takeaway from this was that it was easier to analyse a relatively feasible finite window of statements (context) closely associated with the input statement that can potentially provide causes. Building on this, the focus shifted to improving efficiency in the context retrieval process and automating the analysis to infer causes from it.

Eventually, the use of Retrieval-Augmented Generation (RAG) was explored with Large Language Models (LLMs) and KGs where RAG is a technique that enhances the capabilities of LLMs by incorporating external knowledge sources, such as KGs, for better reasoning with improved ability to draw logical connections, and generate more accurate and relevant reasoning by extracting relevant contextual knowledge [12, 13]. Hence, RAG was used to create a comprehensive understanding of the causes behind the sentiments expressed in input tweets about COVID 19 between March and August in 2020 in this case, without explicitly representing sentiment information at this stage. The use of LLMs allowed for leveraging GenAI's computational capabilities to query for causes behind the information and sentiments in a statement without manual intervention. By integrating KGs with LLMs, the aim was to generate more contextually relevant and accurate explanations for the sentiments observed in the tweets [14, 15].

This research has significant implications for various stakeholders such as government officials and emergency responders. For policymakers, the insights derived from this approach can inform decision-making processes, especially in areas like crisis management and public health. For businesses, understanding causes behind customer sentiments

on social media can drive better marketing strategies and customer engagement. For researchers, the integration of KGs and LLMs offers a novel approach to analyzing and understanding complex sentiment dynamics in real-time social media data [16] through causal reasoning.

In summary, this research aims to address the challenges and opportunities presented by the abundance of unstructured text-based data from social media by leveraging advanced NLP techniques, KGs, and LLMs. The proposed solution involves the integration of RAG, which combines the strengths of retrieval mechanisms and generative models to efficiently extract relevant context from vast datasets and generate accurate, contextually relevant insights. By automating the root cause analysis of real-time social media data, this approach enhances efficiency and enables a proactive response to dynamic situations. The use of RAG allows the system to quickly retrieve pertinent information and generate insightful explanations, significantly improving the ability to identify and address emerging issues. This research represents a step towards harnessing the full potential of text-based data in the age of Big Data, paving the way for more informed and timely decision-making processes.

1.1 Motivation

During the COVID-19 pandemic, a recurring theme in tweets [17] was an increased appreciation for simple things in life. This observation sparked curiosity, leading to attempts to manually understand the reasons behind this sentiment. The process proved extremely challenging due to the scattered and unstructured nature of relevant tweets. For instance, to answer the question, "What caused people to appreciate simple things around them during the COVID-19 pandemic?" sifting through countless tweets was necessary. These tweets were neither posted sequentially nor directly related, making it difficult to piece together a comprehensive answer manually.

After significant manual effort, it was found that people's appreciation for simple things was influenced by interconnected factors [4]. Lockdown measures led to isolation, heightening awareness of immediate surroundings. Restrictions on activities forced individuals to spend more time at home, allowing them to notice and value the simple things

always present [17]. This combination of factors provided comfort and appreciation during challenging times. This manual process highlighted the need for a more efficient solution, underscoring the necessity of developing a method to automatically integrate and analyze this information to uncover deeper insights [8, 2, 18].

As traditional methods of data analysis for causes fall short when dealing with the complexity and speed of social media data [19], the challenge lies in effectively representing and analyzing this unstructured data to derive actionable insights. This is particularly critical in real-time scenarios, such as during emergencies, where timely and accurate information is crucial for decision-making [20].

Here are some of the several challenges that arise in this context:

- **Volume and Velocity of Data:** The immense volume and rapid generation of social media data make it impractical to analyse manually. Automated methods are therefore essential to keep up with the speed at which data is produced [2].

- **Unstructured Nature of Data:** Social media data is inherently unstructured, noisy and varied, containing text that are generally not semantically uniform and with presence of emojis and emoticons. This diversity complicates the extraction and analysis of relevant information [21].

- **Sentiment Analysis Complexity:** Understanding the sentiments expressed in social media posts is challenging due to the context-dependent nature of human emotions. Accurate sentiment analysis requires advanced NLP techniques that can capture these subtleties [3].

- **Contextual Understanding:** To gain meaningful insights, it is necessary to understand the context in which sentiments are expressed. This involves identifying entities and relationships among them within the whole subset of data. Using KGs can help representing them in a structure that is easier to analyse [9].

- **Real-time Processing:** For applications like crisis management, processing data in real-time to reflect changing conditions is crucial and social media text is a vital

source. Delayed responses can result in missed opportunities to mitigate threats or capitalize on current opinions [4].

- **Integration of Knowledge Sources:** Incorporating multiple knowledge sources into the analysis process can enhance perceptions about factors responsible for the scenarios. However, integrating these sources seamlessly using NLP techniques remains a challenge [5, 7].

Despite the extensive utilization of NLP across various industries to analyze large quantities of data, achieving complete automation, where processes are entirely handled by machines without human intervention, remains a challenge [22]. Many NLP applications still require manual interventions for tasks such as fine-tuning models, interpreting complex contexts, and handling edge cases that automated systems struggle to process accurately. For instance, sentiment analysis can measure public opinion [4], real-time entity recognition can identify and categorize key terms in large text corpora [9], and topic modeling can discover hidden themes within unstructured data. However, these processes often need human oversight to perform tasks such as gathering insights into causes behind sentiments, discovering hidden entity relationships and identifying the right subset of data for context [23]. This gap highlights the ongoing need for human expertise to complement and enhance NLP technologies.

A particularly challenging aspect of NLP is causal reasoning, which requires systems to not only understand and interpret data but also identify cause-and-effect relationships within the text. Data representation, a foundational stage in this pipeline, demands technological innovation to integrate generic decision-making principles and domain-specific information reusable across entities. Collaborative efforts between researchers and industry professionals are crucial for significant advancements in this area [6]. Modern challenges include creating reusable, domain-specific repositories that can support robust causal reasoning frameworks [24].

RAG offers a promising solution to enhance the effectiveness of causal reasoning in NLP. By retrieving relevant context from vast datasets, together with LLM, RAG enhances the generative model's ability to produce accurate, contextually relevant insights.

This approach addresses the limitations of traditional NLP systems by providing a richer, more comprehensive understanding of the data.

In the context of causal reasoning, RAG can dynamically retrieve and incorporate pertinent information, enabling the system to better identify and understand cause-and-effect relationships within the text. This leads to more effective and efficient root cause analysis, which is crucial for real-time decision-making and actionable insights.

1.2 Research Objective

The overarching aim of this research was to explore the effectiveness of different methodologies for representing and utilizing KGs in conjunction with LLMs to enhance the extraction and analysis of causal insights from social media data [5]. The research progresses through various stages, experimenting with different approaches to establish a comprehensive and effective framework for causal analysis. The initial focus is on performing root cause analysis of sentiments, as sentiments are valuable indicators of human perceptions behind statements, which can be analyzed to infer underlying causes [25, 8].

To achieve this goal, the research follows a structured approach, beginning with simpler representations of KGs and gradually incorporating more sophisticated techniques involving LLMs as seen in Figure 1.1. The research questions guiding this study are as follows:

RQ1: Can we infer causality by representing social media data as a KG together with sentiment information?

In the initial phase, the research aims to determine the feasibility of extracting causal insights by representing entire tweets as nodes within a Knowledge Graph. The importance of addressing this lies in its direct contribution to the main research objective of performing causal analysis on social media data. Constructing and analyzing a Knowledge Graph enables the visualization and understanding of complex relationships between tweets. By representing tweets as nodes and their interconnections through edges, we can uncover hidden patterns and insights that are not immediately apparent through traditional analysis methods. This approach allows us to identify potential causal links be-

tween different events and sentiments expressed in social media posts, thereby providing a deeper understanding of the dynamics driving public opinion and discourse. Ultimately, this knowledge can inform more effective strategies for monitoring and responding to social media trends and behaviors. Sentiment analysis algorithms are applied to the tweets to identify the emotional tone, which is then added as a property of the nodes. This approach involves:

- Selecting and evaluating several sentiment analysis algorithms to choose the most effective one [4].
- Performing sentiment analysis on tweets and incorporating the results as properties of the nodes within the KG [9].
- Analyzing the resulting KG to identify potential causal relationships between the sentiments expressed in different tweets [11].

RQ2: Can we infer causality by extracting context from KGs represented with entities as nodes together with sentiment information?

Based on the findings from the initial phase, the research evolved to a more refined approach where entities identified within tweets are represented as nodes in the Knowledge Graph. Relationships between these entities are established based on contextual similarities for the identification of more meaningful connections by considering the specific context in which entities appear within each sentence, and sentiments are expressed as weights on the edges connecting the nodes. The importance of addressing this lies in its contribution to a more refined and effective causal analysis of social media data. By constructing KGs where entities are represented as nodes and sentiment information is encoded as edge weights, this approach facilitates a deeper and more detailed understanding of the intricate relationships within the data. This methodology allows for the identification of subtle causal links by examining the interactions between entities and how sentiments modulate these interactions. This ensures that the analysis captures the contextual nuances and emotional dynamics of the data, thereby leading to more accurate and comprehensive insights into the causal relationships inherent in social media interac-

tions. This technical enhancement supports the overall research objective by significantly improving the ability to detect and analyze the underlying causes behind sentiments expressed in social media posts, utilizing advanced graph-based techniques and sentiment analysis. This phase involves:

- Performing Named Entity Recognition (NER) to identify entities within tweets [26].
- Using cosine similarity with a threshold to form relationships between nodes based on contextual similarities [27].
- Assigning sentiments as edge weights to provide a nuanced view of the relationships between entities and the sentiments associated with them [8].
- Analyzing the enhanced KG to uncover more detailed and accurate causal insights [28].

RQ3: Does the integration of KGs and LLMs through RAG (RAG) enhance causal reasoning capabilities on social media data?

The final phase of the research explores the integration of KGs with LLMs using the RAG (RAG) technique. This approach aims to enhance the causal reasoning capabilities by incorporating external knowledge sources into the LLMs, thereby providing more contextually relevant and accurate explanations for the sentiments observed in social media data. The importance of addressing this question lies in its potential to leverage the capabilities of LLMs and dynamic graph databases, RQ3 aims to improve the construction and updating of KGs. This integration allows for better entity relationship identification, context extraction, and real-time updates, providing a more robust and scalable solution for causal analysis.

This method is evaluated against a baseline model which is an independent LLM (GPT-3.5 Turbo LLM) that performs causal reasoning over the entire volume of data. It has been determined through the results of previous RQ that using a specific context that is closely associated with the input statement to reason causes is better with less computation and time and just visualizations from graphical representations as KGs also didn't prove very effective as observed at the end of RQ1. Hence, this RQ explores using

LLMs for context extraction and has been measured against the results of using LLM without context.

This phase involves:

- Using a fine-tuned LLAMA3 model for performing NER to achieve the best representation for the Knowledge Graph [29].
- Integrating KGs with LLMs through RAG to efficiently extract context and infer causes behind statements [13, 12].
- Evaluating the effectiveness of the integrated approach in enhancing causal reasoning capabilities compared to using LLMs without context [30, 31].
- Investigating how the combination of KGs and LLMs can provide deeper and more accurate insights into the causal relationships within social media data [15, 32].

By addressing these research questions, this study aims to develop a robust framework for the automated causal analysis of real-time social media data. The proposed methodologies seek to enhance the understanding of sentiment dynamics and provide valuable insights into their causes for decision-making processes across various domains [5, 22].

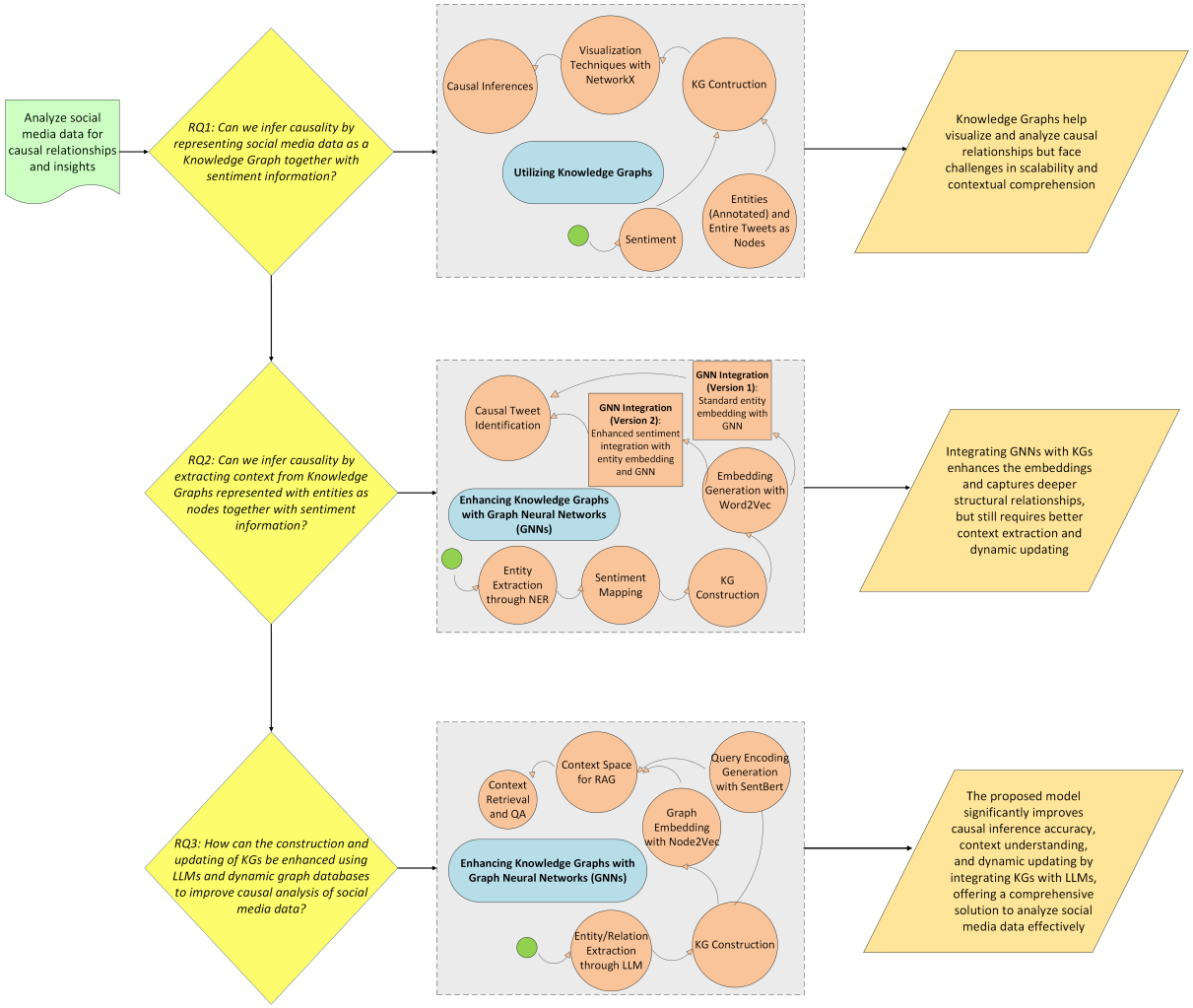

Figure 1.1: Evolution of RQs: Transitioning through Experimentation with Different Ways to Infer Causality

Chapter 2

Related Work

There is a growing interest in computational methods for analyzing dynamic text data, such as social media, has led to significant advancements in data representation and analysis techniques, with KGs and LLMs at the forefront.

2.1 Need for KGs and Sentiment Analysis in Social Media

An essential aspect of analyzing social media data involves constructing KGs. KGs offer a robust framework for structuring and analyzing large volumes of unstructured data, enabling the extraction of meaningful relationships and insights. According to Hogan et al. [24], KGs facilitate data integration, knowledge discovery, and advanced analytics applications such as sentiment analysis and semantic search by organizing information into a graph structure where entities and their interrelations are represented as nodes and edges, respectively. Research by Ehrlinger et al. [33] further emphasizes the importance of KGs in providing a unified schema that integrates diverse data sources, improving data interoperability and accessibility. Additionally, Paulheim et al. [6] discuss methodologies for creating KGs, including entity extraction, relationship identification, and ontology alignment, which are crucial for transforming raw data into structured knowledge representations.

2.2 Sentiment Analysis Techniques

Sentiment analysis in social media has been extensively studied. The paper by Vizcarra et al. [11] explores integrating KGs with sentiment analysis, emphasizing the importance of contextual understanding in social media texts, as discussed also by Khurshid et al.

[8].

2.3 Causal KGs

Constructing causal KGs from varied data sources and enhancing explainability through causal reasoning have been explored in [32, 34, 15]. These studies provide a foundation for the approach used here to creating KGs that predict relations and similarities and offer actionable insights into social media's driving factors. Tan et al. in [32] focused on extracting causal relationships from news articles using linguistic patterns and advanced models like BERT to classify causal sentences, identify causal pairs, and detect cause-effect spans. Utkarshani Jaimini et al. in [34] introduced the CausalKG framework, incorporating interventional and counterfactual reasoning to enhance AI system explainability by representing complex causal relationships in KGs. Seethalakshmi Gopalakrishnan et al. in [15] emphasized using causal reasoning within KGs to uncover the underlying causes of social media sentiments, providing deeper insights for decision-making by systematically identifying and mapping out the cause-and-effect relationships present in social media data, enabling the analysis of how specific events or topics influence public sentiment and behavior. . These studies collectively highlight the importance of contextual similarity in establishing meaningful connections between entities, ensuring accurate causal analysis and offering a richer understanding of data dynamics.

2.4 Causality with LLMs

Exploring the capabilities of LLMs in causal reasoning has been the focus of studies by Kiciman et al. in [35] and Zhijing Jin et al. in [36]. Furthermore, Loureiro et al. in [37] introduces diachronic language models that adapt to the evolving linguistic landscape of social media, enhancing our ability to capture and analyze cause-effect relationships over time.

2.5 Advanced Graph Techniques and NER in Social Media

Lea Dieudonat et al. in [38] explore combining contextual word encodings with KG embeddings, enhancing sentiment analysis through integrated representations. Effective Named Entity Recognition (NER) in social media texts, detailed by Liu et al., informs the preprocessing steps. Additionally, a fine-tuned LLAMA3 model for relation extraction (subject-verb-object triples where verbs are the relations between subject and object entities) demonstrates efficiency in training on data [29].

2.6 RAG Process

The RAG process plays a pivotal role in the methodology by using LLMs and KGs to rank similarities between input text and a list of sentences. This approach provides the top 5 ranked sentences containing causal information, thus reducing the context for further LLM-based cause extraction. Pan et al. in [39] propose frameworks for unifying LLMs and KGs, which was build upon here to enhance predictive power and semantic richness.

For example, let's consider a hypothesis about COVID-19. Suppose an input tweet states, "The number of COVID-19 cases spiked dramatically after the holiday season." The RAG process involves the following steps:

Ranking Similarities: The LLM and KG rank similarities between this input and a list of related sentences. From the COVID-19 dataset, sentences such as:

- "Increased travel during the holiday season led to a surge in COVID-19 cases."
- "Many people ignored social distancing guidelines over the holidays, causing a rise in infections."
- "The holiday season gatherings have been linked to a higher transmission rate of the virus."

Top Ranked Sentences: The RAG process selects the top 5 sentences that contain the most relevant causal information. These sentences help narrow down the context:

- "Increased travel during the holiday season led to a surge in COVID-19 cases."

- "Holiday gatherings without masks contributed significantly to the spike in COVID-19 cases."
- "Post-holiday season saw a peak in COVID-19 cases due to large family gatherings."
- "Ignoring social distancing during holidays resulted in a sharp rise in COVID-19 infections."
- "Travel and parties over the holidays are major factors in the increased COVID-19 case numbers."

LLM-Based Cause Extraction: With the context reduced to these top-ranked sentences, the LLM can more effectively extract the causal relationship. For example, it identifies that "Increased travel" and "Ignoring social distancing during holidays" are key causes of the spike in COVID-19 cases.

2.7 Combining LLMs and KGs

The integration of KGs with LLMs signifies a transformative shift in computational methods. Shirui Pan et al. in [39] outline a comprehensive roadmap for merging LLMs and KGs, which provides the foundation for the work here. This unification harnesses the predictive power of LLMs and the rich semantic structures of KGs, overcoming limitations found in each model when used in isolation. Specifically, LLMs excel in language understanding and generation but often lack the structured, relational knowledge inherent in KGs, leading to potential inaccuracies. On the other hand, KGs provide robust semantic relationships and structured data but do not possess the advanced language processing capabilities of LLMs. By integrating these technologies, our research leverages the language processing strengths of LLMs and the structured knowledge of KGs, by constructing a pipeline that better suits our goal of extracting causal relationships from unstructured social media data, representing these relationships in a KG, and using LLMs to interpret and generate actionable insights.

Chapter 3

Key Concepts

In the realm of data science and artificial intelligence, particularly within the context of analyzing vast amounts of social media data, a robust understanding of several key concepts is essential. This chapter delves into the fundamental techniques and theories that form the backbone of the methodologies employed in this research. The concepts discussed herein, such as KGs [24], ontologies [6], sentiment analysis techniques, and various embedding methods [5], provide the necessary framework for comprehensively understanding and effectively addressing the research questions posed in this study.

The integration of KGs and LLMs is pivotal for extracting and analyzing causal insights from social media data. KGs, which represent entities and their interrelations in a graph format, are instrumental in structuring unstructured data [24]. This structure enables the discovery of hidden relationships and supports advanced reasoning.

Embedding techniques such as Node2Vec, Sentence-BERT, GraphSAGE, and Graph Attention Networks (GAT) play a crucial role in transforming high-dimensional data into more manageable, lower-dimensional representations [27, 40, 41, 42]. These embeddings preserve the semantic relationships within the data, making them suitable for various subsequent tasks, including sentiment analysis and causal inference. By leveraging these embeddings, this research aims to enhance the capability of LLMs to generate contextually relevant and accurate responses [39].

Sentiment analysis techniques, both lexicon-based and machine learning-based, are vital for understanding the emotional tone expressed in social media posts [4, 11], that refers to the underlying sentiment or mood conveyed by the social media posts, whether it is positive, negative, or neutral. They provide insights into public sentiment, which can be a significant indicator of human perceptions and reactions. Understanding these

sentiments is crucial for performing root cause analysis and identifying underlying factors influencing public opinion [4, 11].

Graph Neural Networks (GNNs) are another critical component, designed to operate on graph-structured data. They capture the semantic and relational information inherent in KGs, thus enhancing the analysis of relationships and interactions within the data [43]. GNNs facilitate the embedding of KGs into various Natural Language Processing (NLP) tasks, improving the overall analytical capabilities of the models [44].

Causality in text analysis, particularly within social media data, presents unique challenges due to the unstructured, noisy, and rapidly changing nature of the content [45]. This research integrates KGs with LLMs using Retrieval-Augmented Generation (RAG) to enhance causal analysis [13, 12]. RAG combines the retrieval of relevant information with the generative capabilities of LLMs, thereby improving the quality and relevance of generated text by grounding it in retrieved information [46].

Recent advancements in LLMs, such as GPT-3 and GPT-4, have significantly improved the ability to generate human-like text and provide explanations [47, 48]. These models, when integrated with KGs, offer powerful tools for generating coherent narratives and causality explanations. This integration not only enhances the contextual understanding of the data but also ensures that the generated responses are accurate and relevant [29].

This chapter provides an in-depth exploration of the key concepts essential for the methodologies employed in this research. Understanding these foundational elements is crucial for comprehending the advanced techniques and analyses that will be discussed in the subsequent sections. The methodologies section will build upon these concepts, demonstrating their practical application in answering the research questions addressed in this thesis and achieving the overall research objective.

3.1 Knowledge Graphs

A KG is a powerful data structure designed to represent and store complex relationships between entities in a flexible, schema-less, and relation-oriented manner [24]. KGs consists of nodes and edges, where nodes represent entities and edges represent relationships

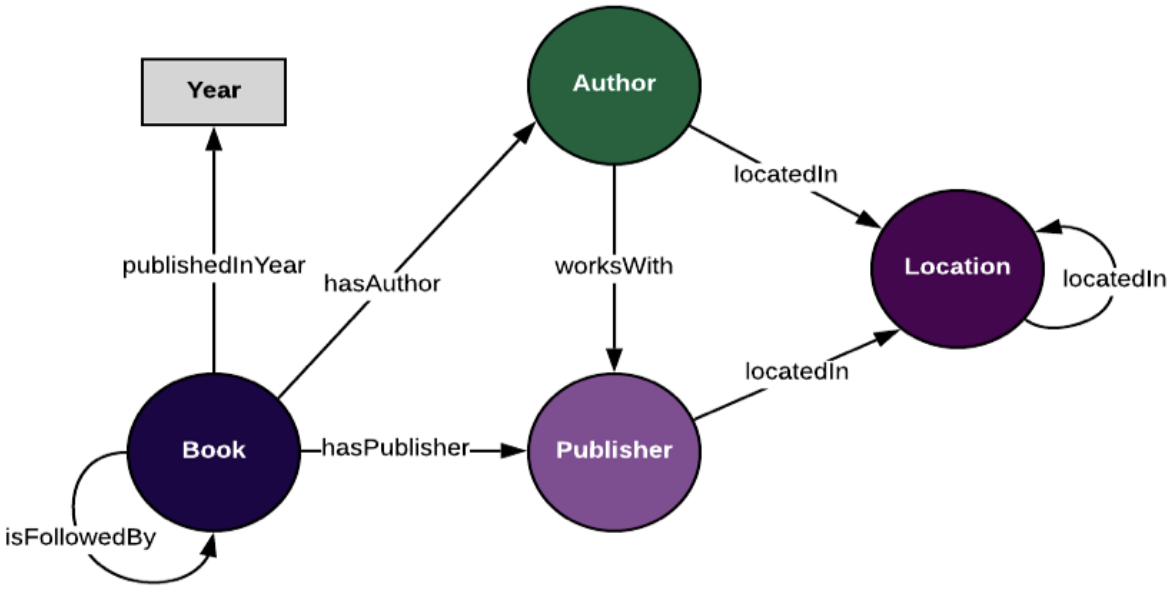

Figure 3.1: Example Representation of a Knowledge Graph from [1]

between them. This structure enables the storage and retrieval of vast amounts of interconnected information, allowing for the discovery of hidden relationships and new insights through deep learning techniques.

Fundamental concept in KGs is a "triple," which consists of a subject, predicate, and object. For example, in the triple (London, is capital of, England), "London" is the subject, "is capital of" is the predicate, and "England" is the object. This simple structure forms the building block for more complex KGs, allowing for the representation of intricate relationships within data [6].

Figure 3.1 from [1] comprises classes like a publisher, an author, and a book. The connections between these entities, along with the publication date, create the contextual links. The entities are represented graphically as nodes and connected by edges representing the relationships between them. A node A and a node B connected by an edge is the smallest KG that can be built and is also called a triple. KGs can take various shapes and sizes, depending on the complexity and nature of the data being represented as discussed in [5].

The process of constructing KGs involves multiple steps, each critical for accurately representing and leveraging the underlying data. This process can be iterative, so that updates can made to accommodate changing contexts over time. The steps include:

Relations Extraction

Relations extraction is the process of identifying and defining relationships between entities in a dataset. These relationships, often expressed as predicates or main verbs, connect entities within the data. For example, in the sentence "Sixty Hollywood musicals were released in 1929," the verb "released in" acts as the predicate connecting "Sixty Hollywood musicals" (subject) and "1929" (object) [16].

Sentence Segmentation

Sentence segmentation involves splitting the text of a document into individual sentences. This step is essential for processing text data, as it delineates where each sentence starts and ends. Rules for sentence segmentation can be based on punctuation marks, user-defined identifiers, or other factors [5].

Entities Extraction

Entities extraction, also known as Named Entity Recognition (NER), involves identifying and classifying entities within a text [9]. These entities can include names of people, organizations, locations, dates, and more. NER is a fundamental technique in Natural Language Processing (NLP) that helps convert unstructured data into a structured form by categorizing key terms within predefined categories. For example, in the sentence "London is the capital of England," NER would identify "London" and "England" as location entities [26].

Data Representation

The final stage in the process is data representation, where the extracted entities and relationships are combined to form a KG. This graph consists of nodes (entities) and edges (relationships) that represent triples. For instance, from the sentences "London is the capital of England" and "Westminster is located in London," we can form the triples (London, is capital of, England) and (Westminster, is located in, London). These triples can then be used to construct a sample KG, visually representing the connections between

entities [15].

Constructing KGs helps achieve the research goal by providing a structured representation of social media data, allowing for efficient retrieval and analysis of relationships. Other techniques explored include network analysis and semantic web technologies, which offer different perspectives on data representation and query optimization.

3.2 Ontologies

Ontologies are structured frameworks that define the categories, concepts, and relationships within a specific domain. They are a subset of linguistics known as semantics, which studies meaning. Ontologies help organize knowledge by categorizing entities and defining their properties and relationships. Typically, they include binary relationships, which describe connections between exactly two entities or concepts, such as "Bear is a mammal" or "Paw part bear" [33].

Ontologies utilize semantic triples, a kind of triple stored in specialized databases called triple stores. These triples are connected whenever they share a common entity, forming a semantic graph. This graph can be used to construct comprehensive databases, enabling efficient querying and retrieval of interconnected data [6].

The process of building ontologies is similar to constructing KGs but focuses on the semantic analysis of text. It involves identifying different types of textual elements based on their meanings. The steps include:

Lexicons

Lexicons are collections of words, terms, and phrases associated with a knowledge domain without context. They represent raw data and are used to filter out unnecessary terms, ensuring that only relevant information is considered [6].

Glossary

A glossary provides single definitions for terms within a lexicon, as defined by expert systems. These definitions are applicable to the knowledge domain associated with the

glossary, ensuring clarity and consistency [5].

Taxonomy

A taxonomy organizes terms by their relationships to one another in a hierarchical structure. Unlike a glossary, a taxonomy lacks detailed definitions but arranges terms to show their relative positions and connections. In NLP, taxonomies are created after lexicons are filtered and glossaries are identified, often represented in forms such as word clouds [33].

Ontology

Ontologies are built from taxonomies and provide a more organized and rule-based structure. They define relationships like hypernymy (words with broader meanings, such as "color" being a hypernym of "red") and can incorporate relationships from multiple overlapping taxonomies to create generalizations across domains [6].

Utilizing ontologies helps achieve the research goal by providing a semantic framework that enhances data integration and retrieval, leading to more meaningful analysis. Other techniques explored include concept maps and thesauri, which offer simpler, less formal structures for organizing knowledge.

3.3 Key Differences between KGs and Ontologies

While KGs and Ontologies share similarities in representing structured knowledge, they differ in key aspects:

- **Knowledge Graphs:** These are focused on representing knowledge in graph form, often derived from graph databases (GDB) that store data using graph architecture. KGs are application-specific, created using ontologies and populated with real data based on domain knowledge [6].

- **Ontologies:** These are generalized data models that define the kinds of things present in a domain and their properties, without including specific individuals.

Ontologies provide definitions of concepts and relationships for a given domain and follow domain rules, making them more abstract and general [6].

Understanding the differences between KGs and Ontologies helps in selecting the appropriate model for specific applications, enhancing the effectiveness of the research methodologies. Other data modeling techniques explored include entity-relationship diagrams and data flow diagrams, which offer different levels of abstraction and detail.

3.4 Decision Making with Text-based Data

Machine Learning techniques that perform reasoning are essential for creating contextual awareness within existing representations. These techniques leverage the abundant computing power available today to reduce the need for manual updates to data. The following parameters are central to decision-making in this research:

- **Domain-specific Research:** Applying KGs and Ontology models to specific domains, focusing on expert systems and personalized representations [5].
- **Generic Knowledge Bases:** Utilizing prior knowledge to build upon new information for NLP models, enhancing their inference capabilities [24].
- **Reasoning on Represented Data:** Employing reasoning techniques for real-time contextual observations on existing trained corpora, improving decision-making processes [6].

This research utilizes KGs to effectively communicate comprehensive sentiment knowledge, guiding domain experts in making informed decisions [33]. Other techniques explored include rule-based systems and expert systems, which provide different approaches to automated reasoning.

3.5 Embedding Techniques

Embedding techniques transform high-dimensional data into lower-dimensional spaces, preserving semantic relationships for various downstream tasks. Key techniques include:

3.5.1 Graph Embedding

Node2Vec

Node2Vec generates node embeddings by simulating random walks on the graph and applying the Skip-Gram model. This technique captures both local and global structural information, making it effective for graph-based tasks. The objective function maximizes the likelihood of observing a node's neighborhood given its embedding, using the softmax function to compute probabilities [27].

$$\max \sum_{u \in V} \sum_{v \in N(u)} \log Pr(v|u)$$

where $Pr(v|u)$ is the probability of node v being a neighbor of node u, computed using the softmax function:

$$Pr(v|u) = \frac{\exp(z_v^\top z_u)}{\sum_{w \in V} \exp(z_w^\top z_u)}$$

Here, z_u and z_v are the embeddings of nodes u and v, respectively.

3.5.2 Text Embedding

Word2Vec

Word2Vec is a widely-used embedding technique that transforms words into continuous vector representations. It uses neural network models to learn word associations from a large corpus of text. There are two main architectures for Word2Vec: Continuous Bag of Words (CBOW) and Skip-Gram [49].

CBOW: Predicts the target word from a given context of surrounding words. It is faster and efficient for large datasets.

Skip-Gram: Predicts surrounding words given a target word. It works well with small amounts of data and represents rare words effectively.

The objective of Word2Vec is to maximize the log probability of the context words given a target word:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log Pr(w_{t+j}|w_t)$$

where T is the total number of words in the corpus, c is the context window size, w_t is the target word, and w_{t+j} are the context words. The probability $Pr(w_{t+j}|w_t)$ is computed using the softmax function:

$$Pr(w_{t+j}|w_t) = \frac{\exp(v_{w_{t+j}}^{\top} v_{w_t})}{\sum_{w=1}^{W}\exp(v_w^{\top} v_{w_t})}$$

Here, v_w is the vector representation of word w, and W is the vocabulary size. This technique captures semantic relationships between words, enabling various NLP tasks such as word similarity, analogy solving, and language modeling [49].

Sentence-BERT

Sentence-BERT is a modification of the BERT (Bidirectional Encoder Representations from Transformers) model designed to generate sentence embeddings. It transforms input sentences into a lower-dimensional vector space where semantically similar sentences are closer together. Sentence-BERT is optimized for sentence-level semantic similarity, making it suitable for tasks such as text clustering, semantic search, and sentence similarity [40].

Using embedding techniques helps achieve the research goal by transforming complex data into manageable representations, enabling efficient analysis and modeling. Other techniques explored include TF-IDF and GloVe, which offer different methods for word and sentence embeddings [50].

3.6 Graph Databases

Graph databases are specialized database management systems designed to store, manage, and query graph-structured data [51]. They excel in handling complex relationships between data points, making them ideal for applications that require the representation and analysis of interconnected data.

3.6.1 Neo4j

Neo4j is a leading graph database platform known for its robust handling of complex relationships within large datasets. It uses a flexible schema and a query language called Cypher, which allows for efficient querying and manipulation of graph data. Neo4j is particularly suited for managing dynamic, heterogeneous data such as social media content, enabling timely analysis and sophisticated queries for deeper insights [20].

3.6.2 Properties of Graph Databases

Graph databases like Neo4j offer several key properties:

- **Schema-less Flexibility:** Graph databases do not require a predefined schema, allowing for the flexible addition of new data and relationships as the graph evolves. This flexibility is crucial for accommodating the dynamic nature of social media data [24].
- **Efficient Relationship Management:** Graph databases are optimized for managing and querying complex relationships between data points, essential for tasks that involve exploring and analyzing interconnected data [20].
- **Temporal Data Handling:** Graph databases can store temporal information on edges, accurately representing event chronology and causal links. This feature is vital for understanding the sequence and timing of events, providing essential insights into causation [43].
- **High Performance and Scalability:** Graph databases are designed to handle large volumes of data and complex queries efficiently, making them suitable for real-time applications that require quick access to and analysis of vast amounts of interconnected data [20].

3.6.3 Use Cases in This Research

In this research, Neo4j is utilized to store and manage the KG constructed from social media data. The graph database facilitates efficient retrieval and analysis of relation-

ships between entities, enabling the exploration of causal dynamics within the data. By leveraging the capabilities of Neo4j, this research aims to uncover deeper insights into the factors driving social media interactions and enhance the overall understanding of causal relationships [24].

Using graph databases helps achieve the research goal by efficiently managing and querying complex relationships within large datasets, enhancing the analysis of interconnected data. Other database technologies explored include relational databases and NoSQL databases, which offer different strengths and limitations for data management.

3.7 Sentiment Analysis Techniques

Sentiment analysis involves using NLP techniques to determine the emotional tone behind words, typically in text form. Various models and algorithms are employed for sentiment analysis, including both machine learning and lexicon-based approaches.

3.7.1 Lexicon-Based Approaches

Lexicon-based approaches rely on pre-defined dictionaries of words associated with specific sentiments. Common lexicon-based tools include:

- **VADER (Valence Aware Dictionary and sEntiment Reasoner):** A lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media [4].
- **SentiWordNet:** An extension of WordNet, where each synset is associated with numerical scores representing positive, negative, and objective sentiments [11].

3.7.2 Machine Learning Approaches

Machine learning approaches for sentiment analysis include:

- **Naive Bayes:** A probabilistic classifier that applies Bayes' theorem with strong (naive) independence assumptions between features [4].

- **Support Vector Machines (SVM):** A supervised learning model that analyzes data for classification and regression analysis [4].

- **Logistic Regression:** A statistical model that uses a logistic function to model a binary dependent variable [11].

The choice of Naive Bayes, Support Vector Machines (SVM), and Logistic Regression for sentiment analysis is based on their proven effectiveness and widespread use in natural language processing tasks. These algorithms are known for their robustness, efficiency, and accuracy in classifying text data, making them suitable for sentiment analysis. Naive Bayes is favored for its simplicity and performance with high-dimensional data, as demonstrated in [4]. SVM is renowned for its ability to handle large feature spaces and its effectiveness in binary classification tasks, as highlighted in [4]. Logistic Regression, on the other hand, is a well-established statistical model that excels in binary classification, providing probabilistic outputs that are easy to interpret, as discussed in [11].

3.7.3 Model Training and Evaluation

The three models (Bernoulli Naive Bayes, Linear SVC, and Logistic Regression) were compared using metrics such as accuracy, precision, recall, and F1-score. The model with the highest performance metrics was selected for further sentiment analysis of the tweet dataset. The steps involved were as follows:

Data Preprocessing

The dataset was preprocessed to clean and normalize the text data. This involved:

- Converting text to lowercase.

- Replacing URLs and user mentions with placeholder tokens.

- Removing non-alphanumeric characters.

- Lemmatizing words to their base forms.

- Replacing emojis with corresponding words based on a predefined emoji dictionary [4].

TF-IDF Vectorization

The preprocessed text data was converted into numerical features using TF-IDF vectorization. This method transforms text into vectors that reflect the importance of words in the documents [21]:

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$$

where $\text{tf}(t, d)$ is the term frequency of term t in document d, and $\text{idf}(t, D)$ is the inverse document frequency of term t in the document set D.

The vectorizer was set to generate n-grams (up to 2-grams) and was limited to a maximum of 500,000 features [21]. The models were then trained on the vectorized data

The models were evaluated on a test dataset using the following metrics:

- **Accuracy:** The proportion of true results among the total number of cases examined.
- **Precision:** The proportion of true positive results among the total number of positive results predicted by the model.
- **Recall:** The proportion of true positive results among the total number of actual positives.
- **F1-Score:** The harmonic mean of precision and recall.

3.7.4 Evaluation Results

The evaluation results for the three models are presented in Table 3.1.

Table 3.1: Results of Sentiment Analysis Models

Model	**Precision**	**F1-Score**
Bernoulli Naive Bayes	0.80	0.80
Linear SVC	0.81	0.82
Logistic Regression	0.82	0.83

Based on these results, the Logistic Regression model was found to be the most efficient, achieving the highest performance metrics. It was therefore selected for the sentiment analysis of the tweet dataset [4].

Employing sentiment analysis techniques helps achieve the research goal by understanding the emotional tone of social media posts, providing insights into public sentiment. Other techniques explored include emotion detection and opinion mining, which offer more granular analysis of text sentiment.

3.8 Graph Neural Networks in NLP

Graph Neural Networks (GNNs) are a class of neural networks designed to operate on graph-structured data. They are particularly effective in tasks involving relationships and interactions within a graph, making them ideal for embedding KGs in NLP tasks [41]. GNNs help capture the semantic and relational information in the graph, which is crucial for the tasks of sentiment and causal analysis in this research.

Incorporating GNNs enhances the representation and analysis of relationships within graph-structured data, thereby achieving the research goal. Other techniques explored include convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which offer different approaches to processing sequential and grid-like data. However, CNNs are primarily designed for spatial data and excel in image recognition tasks, while RNNs are tailored for temporal data and are effective in handling sequences, such as time-series analysis or language modeling.

GNNs are specifically designed to work with graph data, allowing them to propagate information across nodes and edges effectively. This capability is crucial for capturing the complex relationships and hierarchies present in graph-structured data. GNNs leverage

the power of message passing, where nodes aggregate information from their neighbors iteratively, leading to a more comprehensive understanding of the graph's global structure. This makes GNNs the most favored technique for our research, as they provide a robust framework for modeling and analyzing the rich, interconnected data inherent in social media networks and other graph-based applications.

3.9 Causality in Text Analysis

Causality in text analysis refers to identifying and understanding the cause-and-effect relationships expressed in natural language. This involves detecting patterns and inferring causal links between events, actions, and sentiments described in the text. Effective causal analysis can provide deeper insights into the reasons behind sentiments and opinions expressed in social media data [52, 32].

Causality in Social Media Text

Analyzing causality in social media text involves identifying and understanding the cause-and-effect relationships expressed in user-generated content. This section delves into the methodologies and challenges associated with extracting causal relationships from social media platforms, using a combination of KGs and LLMs.

Challenges in Causal Analysis of Social Media Data

Social media data presents several unique challenges for causal analysis:

- **Unstructured and Noisy Data:** Social media posts are often unstructured, containing slang, abbreviations, and noise such as hashtags and emojis.
- **Temporal Dynamics:** The context and relevance of social media posts change rapidly, requiring methods that can handle temporal data effectively [37].
- **High Volume and Velocity:** The sheer volume and speed of data generation on social media necessitate scalable and efficient analytical techniques.

Methods for Causal Analysis

Several methodologies are employed to tackle these challenges and extract causal relationships from social media data:

- **Temporal Analysis:** Examining the sequence and timing of events to identify potential causal links [28].
- **Sentiment Analysis:** Determining the sentiment expressed in posts to understand how sentiments influence and are influenced by events [4].
- **Entity Recognition:** Identifying and linking entities mentioned in posts to explore relationships between them [26].
- **Graph-Based Methods:** Utilizing KGs to represent and analyze relationships between entities and events [7, 16].

Integration of Knowledge Graphs and LLMs

This research integrates KGs with LLMs to enhance causal analysis of social media data. The integration leverages the structured representation of KGs and the contextual understanding of LLMs to uncover deeper causal insights [39].

Knowledge Graph Construction: The construction of the KG involves identifying entities and relationships from social media posts using Named Entity Recognition (NER) and relation extraction techniques [9]. The KG represents entities as nodes and relationships as edges, with temporal information stored as edge properties [52].

Embedding and Encoding: The KG is embedded using techniques such as Node2Vec and Sentence-BERT, generating vector representations that capture the semantic and relational context of the nodes and edges [40].

Retrieval-Augmented Generation (RAG): The RAG process combines the retrieval of relevant information from the KG with the generative capabilities of LLMs. The encoded query is used to retrieve contextually similar subgraphs from the KG. The LLM then generates a response based on the retrieved context, providing a detailed and contextually relevant causal explanation [12, 30].

Advantages of the Integrated Approach

The integration of KGs and LLMs offers several advantages for causal analysis:

- **Enhanced Contextual Understanding:** Combining structured knowledge in KGs with the contextual understanding of LLMs leads to more accurate and relevant causal explanations [53].
- **Improved Accuracy:** KGs provide a structured and verified context, reducing the risk of generating inaccurate or misleading responses [31].
- **Scalability:** The approach is scalable and capable of handling large volumes of social media data, providing real-time causal insights.
- **Temporal Relevance:** The integration allows for the inclusion of temporal information, ensuring that generated explanations respect the chronological order of events [37].

In summary, integrating KGs and LLMs provides a robust framework for performing causal analysis on social media text, offering deeper contextual understanding, improved accuracy, and scalability [35, 54].

Applying causality analysis techniques helps achieve the research goal by identifying cause-and-effect relationships within social media data, offering deeper insights into underlying factors. Other techniques explored include statistical methods for causal inference and time series analysis, which provide alternative approaches to causality detection.

3.10 Recent Advances in LLMs for Text Generation

LLMs like GPT-3 and GPT-4 have revolutionized the field of natural language generation and understanding. These models leverage vast amounts of text data to generate human-like text, answer questions, and provide explanations. Recent advancements in LLMs have shown promise in generating causality explanations by understanding context and generating coherent narratives [48].

3.10.1 LLM Enhanced KG

Using LLMs for KG generation have significantly enhanced it's capabilities. By leveraging the vast contextual understanding and text generation abilities of LLMs, KGs can be enriched with more nuanced and semantically rich data [5].

LLMs can assist in the following ways:

- **Entity Recognition and Linking:** LLMs can identify entities within a text and link them to existing entities in the KG, thereby expanding the graph with new information [9].
- **Relation Extraction:** Using their understanding of natural language, LLMs can extract complex relationships between entities that might be missed by traditional NLP techniques [53, 29].
- **Contextual Enrichment:** LLMs can provide contextual information and background knowledge that enriches the nodes and edges in a KG, making the graph more informative and useful for downstream applications [39].
- **Data Augmentation:** LLMs can generate additional data based on existing entities and relationships in the KG, thus expanding the graph in a meaningful way [2].

The usage of LLMs with KGs allows for more dynamic and comprehensive representations of knowledge, facilitating advanced reasoning and decision-making processes [13, 12].

3.10.2 KG Enhanced LLM

KGs can enhance LLMs by providing structured and relational context that helps LLMs generate more accurate and contextually relevant responses.

KGs enhance LLMs through:

- **Contextual Information:** KGs provide structured context to the LLMs, enabling them to generate responses that are consistent with the knowledge encoded in the graph [7, 19].

- **Disambiguation:** With the help of KGs, LLMs can better disambiguate entities and relations in the text, leading to more precise and accurate text generation [16].

- **Consistency and Coherence:** KGs help maintain consistency and coherence in the responses generated by LLMs by ensuring that the generated text adheres to the relationships and constraints defined in the graph [22].

- **Query Expansion:** LLMs can use the relational data in KGs to expand and refine queries, leading to more comprehensive and relevant answers [13].

By incorporating KG data, LLMs can generate outputs that are not only syntactically and semantically correct but also aligned with the structured knowledge, improving the overall quality of text generation [6, 33].

3.10.3 Retrieval Augmented Generation (RAG)

RAG is a technique that combines the retrieval of relevant documents or information with the generative capabilities of LLMs. This approach aims to enhance the quality and relevance of generated text by grounding it in retrieved information [12, 30].
Key components of RAG include:

- **Retrieval Mechanism:** This component involves retrieving relevant information from a large corpus or database based on the input query. The retrieval system uses techniques such as TF-IDF, BM25, or dense embeddings to find the most relevant documents or passages [21, 49].

- **Generative Model:** The retrieved information is then used to condition the generative model (LLM), which generates a response that incorporates the retrieved context [3].

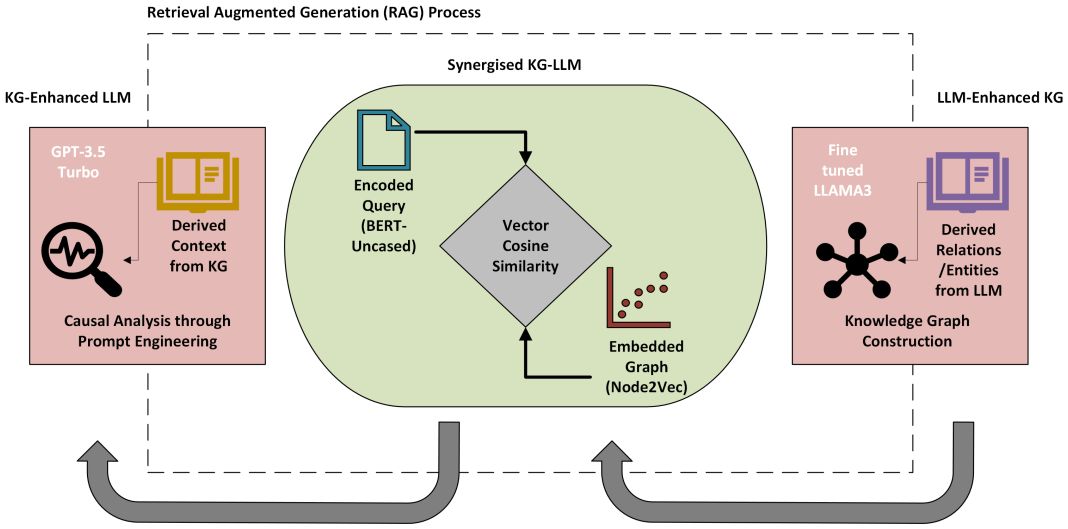

Figure 3.2: Combining KG and LLM through Retrieval Augmented Generation (RAG)

- **Fusion Strategy:** The final output is generated by fusing the retrieved information with the generative capabilities of the LLM. This can involve concatenating the retrieved text with the input query or using more sophisticated techniques to integrate the information [12, 27].

RAG models leverage the strengths of both retrieval and generation, providing several benefits:

- **Improved Relevance:** By grounding the generated text in retrieved information, RAG ensures that the responses are more relevant and factually accurate [30, 12].
- **Enhanced Coherence:** The retrieved context helps maintain coherence in the generated text, especially for complex queries requiring detailed information [46].
- **Knowledge Integration:** RAG allows for the seamless integration of external knowledge sources, making the generated responses richer and more informative [13].

The RAG approach is particularly useful in applications such as question answering, summarization, and dialogue systems, where both relevance and generation quality are critical [31, 55].

Integrating recent advances in LLMs with techniques like RAG helps achieve the research goal by improving the relevance and accuracy of generated text, enhancing the understanding of social media data. Other techniques explored include neural machine translation and summarization, which offer different applications of LLMs in natural language processing.

For this research, KG has been experimented with as a potential source for context retrieval which had been enhanced through LLM that ultimately would pave way for enhanced reasoning capabilities of a generative AI (GPT-3.5 Turbo LLM) as seen in Figure 3.2.

Chapter 4

RQ1: Can we infer causality by representing social media data as a KG together with sentiment information?

The first research question (RQ1) investigates whether causal insights can be gathered from KGs where each tweet is represented as a node with sentiments as node properties. The solution to this question involves exploring various techniques for building and analyzing KGs, comparing them with alternative methods, and identifying the most effective approach.

Tailoring the Solution to Address RQ1

To address RQ1, the initial step is to construct a KG where nodes represent entire tweets, and edges capture the relationships between these tweets based on sentiment analysis. The primary objective is to determine if this structure can reveal causal insights by examining the connections and sentiment properties of the nodes.

Various methods could potentially address RQ1, including text classification and clustering, topic modeling, and dependency parsing combined with event extraction. Text classification categorizes tweets based on content, and clustering groups similar tweets, identifying patterns and trends but often failing to capture the relationships or underlying causality between tweets. Topic modeling techniques, like Latent Dirichlet Allocation (LDA) [56], identify themes within tweets but do not elucidate the causal connections between them. Dependency parsing identifies sentence structure, and event extraction focuses on events and their participants, offering detailed insights but struggling with scalability for large datasets and complexity in implementation. These alternative methods have limitations in capturing the relational dynamics and underlying causes within social media discourse. Therefore, constructing and analyzing a KG (KG) using a com-

bination of techniques provides a more comprehensive approach. This method captures both the structural and relational information within tweets, offering deeper insights into causal relationships and sentiment propagation, making it the most effective solution for addressing RQ1 based on current research.

4.1 Solution Statement

The proposed solution for RQ1 involves constructing and analyzing a KG to capture and visualize causal relationships within tweets. This model is designed to represent each tweet as a node, with edges representing relationships based on entity extraction and sentiment analysis. The core components of the solution are as follows:

1. **Data Preprocessing**: Converting raw data into a structured format suitable for analysis.
2. **Entity Extraction**: Identifying named entities within each tweet and categorizing them.
3. **KG Construction**: Building a graph with nodes representing tweets and edges representing relationships and sentiments.
4. **Visualization and Analysis**: The KG is visualized and multiple analysis techniques are employed to gain insights from the KG.

This comprehensive model captures the structural and relational information within the tweets, enabling the identification of causal relationships and sentiment propagation patterns.

4.2 Dataset

The dataset used for the first research question (RQ1) involves Named Entity Recognition (NER). The objective of RQ1 is to gather causal insights from KGs where each tweet is represented as a node with sentiments as node properties. The dataset for this phase is an annotated corpus specifically designed for NER tasks [57].

Annotated Corpus for Named Entity Recognition

Task: Named Entity Recognition (NER) is a task of categorizing entities in a text into categories like names of persons, locations, organizations, etc.

Table 4.1: Sample Subset of the Annotated Data with Named Entity Tags and Sentiments

Sentence #	Sentence	POS	Tag	Sentiment
1	Thousands of demonstrators have marched through London to protest the war in Iraq and demand the withdrawal of British troops from that country .	['NNS', 'IN', 'NNS', 'VBP', 'VBN', 'IN', 'NNP', 'TO', 'VB', 'DT', 'NN', 'IN', 'NNP', 'CC', 'VB', 'DT', 'NN', 'IN', 'JJ', 'NNS', 'IN', 'DT', 'NN', '.']	['O', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'O', 'O', 'O', 'O', 'B-gpe', 'O', 'O', 'O', 'O', 'O']	neutral
2	Families of soldiers killed in the conflict joined the protesters who carried banners with such slogans as " Bush Number One Terrorist " and " Stop the Bombings . "	['NNS', 'IN', 'NNS', 'VBN', 'IN', 'DT', 'NN', 'VBD', 'DT', 'NNS', 'WP', 'VBD', 'NNS', 'IN', 'JJ', 'NNS', 'IN', '``', 'NNP', 'NN', 'CD', 'NN', '``', 'CC', '``', 'VB', 'DT', 'NNS', '.', '``']	['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-per', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']	negative
3	They marched from the Houses of Parliament to a rally in Hyde Park .	['PRP', 'VBD', 'IN', 'DT', 'NNS', 'IN', 'NN', 'TO', 'DT', 'NN', 'IN', 'NNP', 'NNP', '.']	['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'I-geo', 'O']	neutral
4	Police put the number of marchers at 10,000 while organizers claimed it was 1,00,000 .	['NNS', 'VBD', 'DT', 'NN', 'IN', 'NNS', 'IN', 'CD', 'IN', 'NNS', 'VBD', 'PRP', 'VBD', 'CD', '.']	['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']	neutral
5	The protest comes on the eve of the annual conference of Britain 's ruling Labor Party in the southern English seaside resort of Brighton .	['DT', 'NN', 'VBZ', 'IN', 'DT', 'NN', 'IN', 'DT', 'JJ', 'NN', 'IN', 'NNP', 'POS', 'VBG', 'NNP', 'NNP', 'IN', 'DT', 'JJ', 'JJ', 'NN', 'NN', 'IN', 'NNP', '.']	['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'O', 'O', 'B-org', 'I-org', 'O', 'O', 'O', 'B-gpe', 'O', 'O', 'O', 'B-geo', 'O']	neutral
6	The party is divided over Britain 's participation in the Iraq conflict and the continued deployment of 8,500 British troops in that country .	['DT', 'NN', 'VBZ', 'VBN', 'IN', 'NNP', 'POS', 'NN', 'IN', 'DT', 'NNP', 'NN', 'CC', 'DT', 'JJ', 'NN', 'IN', 'CD', 'JJ', 'NNS', 'IN', 'DT', 'NN', '.']	['O', 'O', 'O', 'O', 'O', 'B-gpe', 'O', 'O', 'O', 'O', 'B-geo', 'O', 'O', 'O', 'O', 'O', 'O', 'B-gpe', 'O', 'O', 'O', 'O', 'O']	neutral
7	The London march came ahead of anti-war protests today in other cities , including Rome , Paris , and Madrid .	['DT', 'NNP', 'NN', 'VBD', 'RB', 'IN', 'JJ', 'NNS', 'NN', 'IN', 'JJ', 'NNS', ',', 'VBG', 'NNP', ',', 'NNP', ',', 'CC', 'NNP', '.']	['O', 'B-geo', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'O', 'B-geo', 'O', 'O', 'B-geo', 'O']	negative
8	The International Atomic Energy Agency is to hold second day of talks in Vienna Wednesday on how to respond to Iran 's resumption of low-level uranium conversion .	['DT', 'NNP', 'NNP', 'NNP', 'NNP', 'VBZ', 'TO', 'VB', 'JJ', 'NN', 'IN', 'NNS', 'IN', 'NNP', 'NNP', 'IN', 'WRB', 'TO', 'VB', 'TO', 'NNP', 'POS', 'NN', 'IN', 'JJ', 'NN', 'NN', '.']	['O', 'B-org', 'I-org', 'I-org', 'I-org', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'B-tim', 'O', 'O', 'O', 'O', 'O', 'B-gpe', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']	neutral
9	Iran this week restarted parts of the conversion process at its Isfahan nuclear plant .	['NNP', 'DT', 'NN', 'VBD', 'NNS', 'IN', 'DT', 'NN', 'NN', 'IN', 'PRP$', 'NNP', 'JJ', 'NN', '.']	['B-gpe', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'O', 'O', 'O']	neutral
10	Iranian officials say they expect to get access to sealed sensitive parts of the plant Wednesday , after an IAEA surveillance system begins functioning .	['JJ', 'NNS', 'VBP', 'PRP', 'VBP', 'TO', 'VB', 'NN', 'TO', 'JJ', 'JJ', 'NNS', 'IN', 'DT', 'NN', 'NNP', ',', 'IN', 'DT', 'NNP', 'NN', 'NN', 'VBZ', 'VBG', '.']	['B-gpe', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-tim', 'O', 'O', 'O', 'O', 'B-org', 'O', 'O', 'O', 'O', 'O']	positive

Dataset: As seen in Table 4.1, each row in the structure is a complete sentence, a list of POS tags for each word in the sentence, and a list of NER tags for each word in the sentence. The dataset was taken from the Annotated Corpus for Named Entity Recognition by Abhinav Walia in [57] and processed to enhance its features.

Dataset Fields

- **sentence**:
  - **Description**: Each row contains a complete sentence from the corpus.
  - **Type**: String
  - **Importance**: This field is essential for understanding the context in which the entities and POS tags appear. It provides the primary text for analysis.

- **pos**:
  - **Description**: List of POS tags for each word in the sentence.
  - **Type**: List of strings
  - **Importance**: POS tags are crucial for syntactic analysis and provide additional context for understanding the role of each word in the sentence.
- **tag**:
  - **Description**: List of NER tags for each word in the sentence.
  - **Type**: List of strings
  - **Importance**: NER tags categorize each word into predefined entity types, which are critical for entity recognition tasks. These tags help in identifying and classifying entities within the text.

Essential Information About Entities

- **geo**: Geographical Entity
- **org**: Organization
- **per**: Person
- **gpe**: Geopolitical Entity
- **tim**: Time indicator
- **art**: Artifact
- **eve**: Event
- **nat**: Natural Phenomenon

Provenance

- **Sources**: The dataset originates from the Annotated Corpus for Named Entity Recognition by Abhinav Walia [57], which uses the GMB (Groningen Meaning Bank) corpus for entity classification.
- **Collection Methodology**: The dataset was annotated using the GMB corpus, involving meticulous annotation of text to identify and categorize named entities. The annotations include both POS tags and NE tags for each word in the sentences.

Usage and Applications

The dataset is valuable for training and evaluating NER models. By leveraging the detailed annotations provided in the dataset, researchers can gain insights into:

- **Named Entity Recognition (NER)**: Training and evaluating models to accurately identify and categorize named entities within text.
- **Part-of-Speech Tagging (POS)**: Understanding the syntactic structure of sentences by analyzing the POS tags associated with each word.
- **Text Analysis**: Extracting meaningful entities and understanding their roles and relationships within the text.
- **Information Extraction**: Identifying and categorizing relevant information from large text corpora for various applications, such as knowledge base construction and content analysis.

Relevance to RQ1

This dataset is ideal for addressing RQ1 because it provides a structured way to extract entities and their relationships from text, which is crucial for constructing KGs. By utilizing the POS and NER tags, we can accurately identify and categorize entities within each sentence. This allows us to represent each tweet as a node in the KG, with edges

representing the relationships between entities. The sentiments associated with each entity further enhance the graph by adding emotional context to the nodes and edges. The combination of entity recognition and sentiment analysis facilitates a deeper understanding of the causal relationships between different elements in the text, making this dataset highly suitable for the research objectives of RQ1.

To demonstrate this, the dataset was processed using the following approach:

- **Data Extraction**: Sentences were extracted along with their POS and NER tags.
- **Entity Identification**: Entities within each sentence were identified and categorized based on their NER tags.
- **Sentiment Association**: Sentiments were associated with each identified entity, adding a layer of emotional context.
- **Graph Construction**: A KG was constructed where each node represented a tweet, and edges represented relationships between entities. Sentiments were used as node properties and edge colors to visualize the emotional tone of the relationships.
- **Visualization and Analysis**: The graph was visualized to identify connected components, and Node2Vec embeddings were generated to capture the structural and relational information within the graph.

This structured approach ensures that the dataset is effectively utilized to address RQ1, providing a comprehensive framework for understanding the causal relationships between entities and sentiments within the text.

4.3 Methodology

The processes discussed in the solution statement are explained in detail here. Figure 4.1 represents the workflow of the complete solution to RQ1.

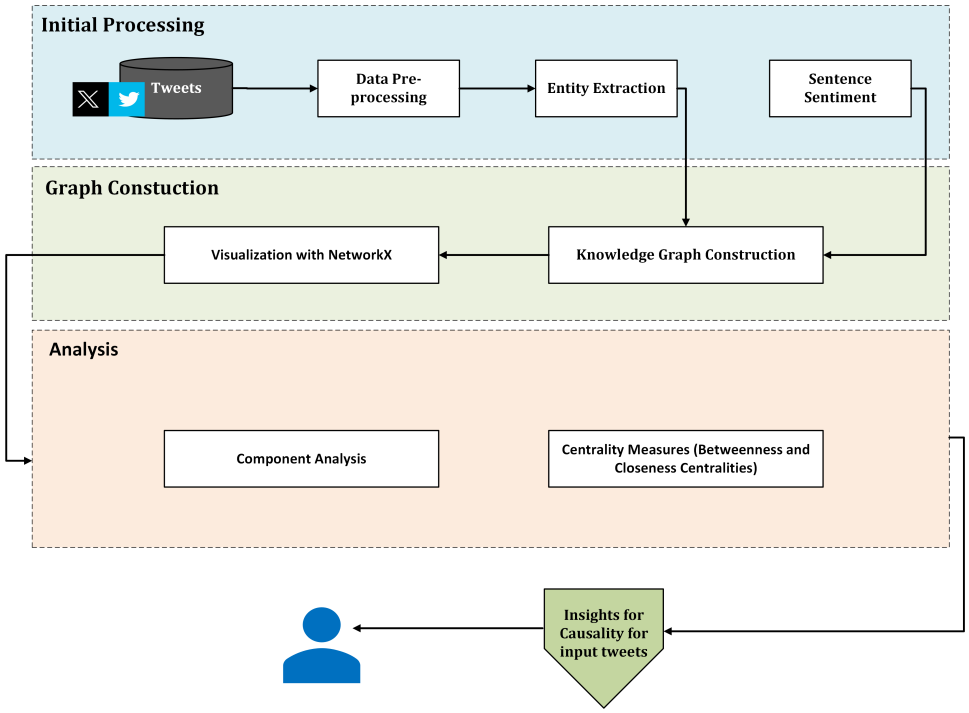

Figure 4.1: RQ1: Methodology for Graph Construction, Visualization and Analysis for Causality

4.3.1 Data Preprocessing

The dataset used for this study is a CSV file containing sentences, Part-of-Speech (POS) tags, and Named Entity Recognition (NER) tags. The data preprocessing phase involves converting the POS and NER tags from string representations to list representations. This conversion ensures that the tags can be used for further analysis.

```
# Pseudo-code for Data Preprocessing
Load CSV data into DataFrame
Convert POS and NER tags from strings to lists
```

4.3.2 Entity Extraction

Once the data is preprocessed, the next step is to extract entities and their types from each sentence. This is achieved by iterating over each word in the sentence along with its POS and NER tags. Named entities are identified based on the NER tags, and they are categorized into different entity types (e.g., geographical entities, organizations, persons).

The process involves initializing variables to track current entities and their types,

iterating through each word in the sentence, and appending entities to lists when they are identified.

```
# Pseudo-code for Entity Extraction
Initialize entity tracking variables
For each word in the sentence:
    If the word is a named entity:
        Append the entity and its type to lists
    Else:
        Reset entity tracking variables
```

4.3.3 KG Construction and Visualization

The next step is to construct the KG and visualize using NetworkX. Nodes in the graph represent entire tweets, and edges capture the relationships between these tweets based on the extracted entities and their sentiment properties. Sentiment analysis is performed to assign sentiment values (positive, neutral, or negative) to the edges, which are then used for further analysis.

The graph is created as a directed graph where each edge has properties such as the relationship type, the original tweet text, and the sentiment. This allows for a detailed representation of the interactions and sentiments between tweets.

```
# Pseudo-code for KG Construction
Initialize directed graph
For each tweet:
    Add node representing the tweet
    For each entity in the tweet:
        Add edge with relationship and sentiment properties
```

To gain insights from the KG, various visualization techniques are employed:

- **Node Visualization**: Nodes are visualized to understand their significance and relationships within the graph.

- **Edge Visualization**: Edges are colored according to sentiment values, illustrating the flow and interaction of sentiments within the network.

The visualization process includes creating scatter plots, coloring nodes based on their sentiment values, and annotating plots with node labels. These visualizations facilitate the calculation and analysis of centrality measures, such as degree centrality, betweenness centrality, and closeness centrality, providing deeper insights into the structure and dynamics of the graph. This approach helps in identifying influential tweets and understanding the distribution of sentiments across the network.

4.4 Results and Analysis

The results presented here are based on a sample subset of the entire dataset to provide an idea of the analysis performed. This subset allows us to demonstrate the methodology and the types of insights that can be derived from the data. First, the results of component analysis and KG contruction are presented, which are followed by analysing them along with centrality measures for insights into causes.

Connected Components Analysis

In the analysis of the KG, several connected components were identified. These components highlight clusters of entities and their relationships within the dataset, revealing potential groupings or communities that share common features. Below are the identified components along with some insights into their significance:

- **Component 1:** This component includes entities such as Labor Party, Madrid, London, Britain, British, Rome, Prophet Muhammad, English, Iraq, Paris, Brighton, and Islam. The presence of both geopolitical entities and significant terms related to religious and cultural discussions suggests this component represents discussions involving European and Middle Eastern geopolitical relations and cultural interactions.

- **Component 2:** Containing Bush, this component is relatively isolated, indicating a specific context where Bush is mentioned without extensive connections to other entities.

- **Component 3:** This component focuses on Hyde Park, indicating localized discussions or events centered around this location.

- **Component 4:** This is a large and complex component including entities such as Royal-Dutch Shell, Wednesday, Anbar, Pakistani, Thomas Horbach, Iran, Isfahan, German, President Mahmoud Ahmadinejad, Mogadishu, Kani Wam, Afghanistan, Taleban, European, Bedford, Malik Faridullah Khan, Delta State, Bedfordshire, Baghdad, Mosul, European Union, Bilfinger Berger, International Atomic Energy Agency, Iranian, American, Bayelsa State, al Qaida in Iraq, IAEA, South Waziristan, Omar Khayam, U.S., Tehran, Nuclear Non-Proliferation, Iraqi, Sunni, U.N. Security Council, Sunday, Tuesday, and Vienna. This component likely represents a broad range of geopolitical interactions, with nodes connected by various events and relationships.

- **Component 5:** This includes entities such as Germans, Delta, Nigerian, and Nigeria, indicating a focus on interactions between German and Nigerian entities.

- **Component 6:** Focuses on Niger Delta, indicating localized discussions or events specific to this region.

- **Component 7:** Including Somalia and President Abdullahi Yusuf Ahmad, this component likely represents discussions around Somali politics and conflicts.

- **Component 8:** Focuses on Somali, indicating discussions related to Somali entities.

- **Component 9:** Includes Kurdish and Sunni Arab, highlighting discussions around ethnic and religious groups in the Middle East.

- **Component 10:** This includes Egyptian, Saturday, Muslim Brotherhood, indicating a focus on Egyptian politics and the activities of the Muslim Brotherhood.

- **Component 11:** Contains Alexandria and Friday, suggesting discussions specific to Alexandria and possibly events occurring on Fridays.
- **Component 12:** Includes Brotherhood, focusing on discussions related to the Muslim Brotherhood.
- **Component 13:** Includes Pakistan and North West Frontier Province, indicating discussions specific to Pakistani regions.
- **Component 14:** Contains Thursday and Mutahida Majlis-e-Amal, indicating discussions around specific events or entities related to Thursday.
- **Component 15:** This component includes British Home Office, Khayam, AP, indicating discussions involving British authorities and related individuals.
- **Component 16:** Focuses on Associated Press and Home Office, indicating interactions involving these entities.

These components provide a clear overview of how different entities are connected and can help in understanding the structure and focus of discussions within the dataset.

KG Visualization with Sentiment Information

As seen in Figure 4.1, the KG incorporates sentiment information, represented by the color of the edges. Green edges indicate positive sentiment, red edges indicate negative sentiment, and gray edges indicate neutral sentiment. The nodes represent statements, with their ID serving as the node name. The outermost nodes are the entities identified from the dataset.

This visualization highlights the relationships between entities and the sentiments associated with these relationships, providing insights into the nature of discussions within the dataset.

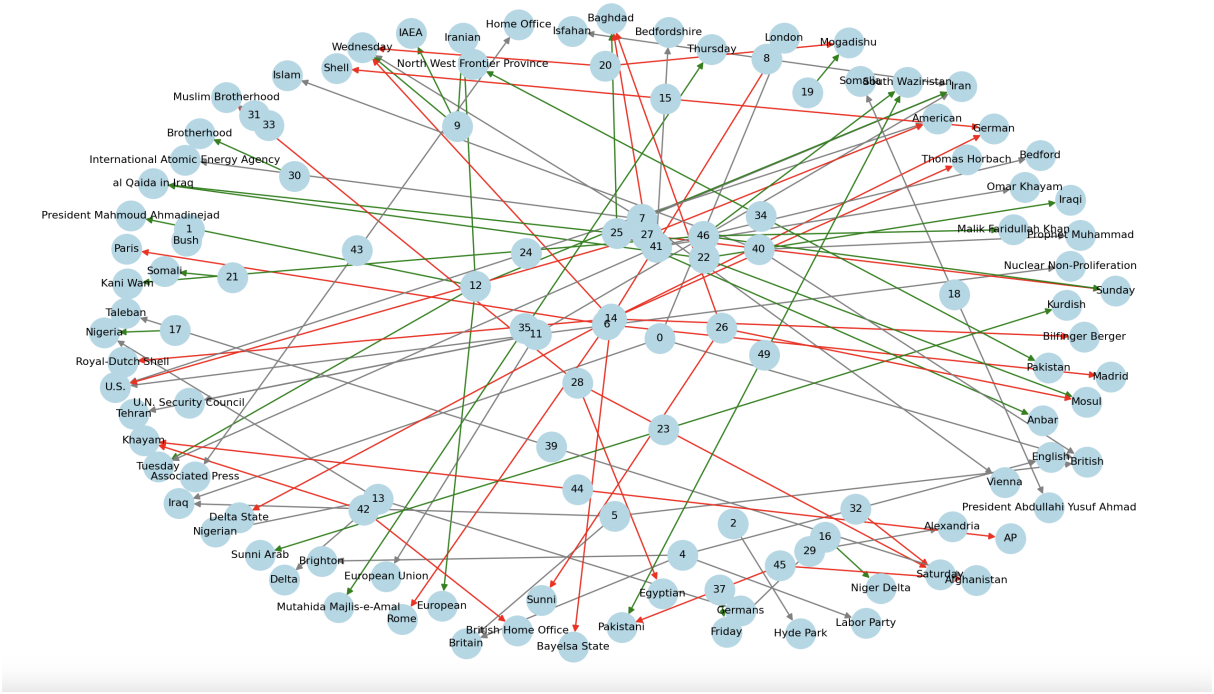

Figure 4.2: RQ1: Visual Map of Tweets Connected by Emotional Tone Using a KG with Sentiment Information

Inferences for Causality

Consider the tweet: "Thousands of demonstrators have marched through London to protest the war in Iraq and demand the withdrawal of British troops from that country."

KG Analysis

Entities Identified: London, Iraq, British troops

Sentiment: Neutral

Relationships: This tweet connects entities such as London, Iraq, and British troops with relationships indicating protest and demand for withdrawal.

Centrality Measures Analysis

The centrality measures indicate the importance of nodes within the graph [6]. High degree centrality nodes are those that are connected to many other nodes, suggesting that they are influential tweets. Betweenness centrality measures the extent to which

a node lies on the shortest paths between other nodes, indicating its role in connecting different parts of the graph. Closeness centrality measures how close a node is to all other nodes in the graph, indicating its potential for quickly spreading information.

Betweenness Centrality: High Betweenness Nodes are Node 1 (Bush), Node 9 (Islam), Node 12 (Royal-Dutch Shell) – These nodes act as crucial connectors, indicating centrality in various geopolitical discussions.

Closeness Centrality: High Closeness Nodes are Node 8 (London), Node 0 (British Home Office), Node 46 (Vienna) – These nodes are quickly accessible within the network, highlighting their influence and central role in discussions.

Component Analysis

Connected Components: The component includes entities like London, Iraq, and British troops, indicating discussions focused on geopolitical protests and military actions.

Entity Groupings: The presence of geopolitical and protest-related entities shows a strong focus on public demonstrations against military actions.

Final Insights

The analysis of the above results reveals several insights. The KG shows a clear connection between London, Iraq, and British troops, with relationships indicating protest and demands for withdrawal. The sentiment analysis tags the tweet as neutral, suggesting that it is a factual report of events without strong emotional undertones. The high betweenness centrality of nodes like Bush and Islam indicates that discussions around these entities are pivotal in the network, acting as bridges connecting various topics. The high closeness centrality of nodes like London and the British Home Office suggests that these entities are central to the network and quickly accessible, highlighting their importance in the discourse. The connected components analysis shows that discussions around this tweet are part of a larger conversation involving geopolitical protests and military actions, with London and Iraq being central to these discussions.

This comprehensive analysis demonstrates how the tailored solution effectively ad-

dresses RQ1 by constructing and analyzing a KG to uncover causal relationships within tweets. By examining the relationships, centrality measures, and connected components, we gain a deeper understanding of the underlying dynamics and influence patterns within the dataset. This approach provides valuable insights for policymakers, researchers, and the public, enabling more informed decision-making and targeted strategies to address the issues discussed in social media.

4.5 Limitations and Transitioning to RQ2

The main disadvantage of this method in analyzing causes from inferred insights is its reliance on explicit relationships and sentiment analysis, missing the underlying nuances and contextual subtleties in social media discourse. The causality inferred here is based on direct connections and sentiment, which can be limiting and not always reflective of true underlying causes. This approach may lead to oversimplified conclusions and overlook significant indirect influences. While the KG maps out direct connections and sentiment flows, it may miss deeper, implicit causal relationships and the influence of indirect or latent factors. Additionally, it may struggle with complex, multifaceted issues where causality involves a web of interacting elements and long-term influences. These limitations underscore the need for more advanced techniques to handle the intricate and dynamic nature of social media data, leading to the investigation addressed in RQ2, focusing on enhancing causal analysis by integrating more sophisticated models and methodologies.

To address these limitations, the construction of the KG can be improved in several ways:

- **Enhanced Entity Disambiguation:** Implementing more sophisticated algorithms for entity recognition and disambiguation can help reduce confusion and improve the accuracy of connections.
- **Contextual Relationships:** Including more contextual information in the relationships can provide a deeper understanding of the connections and improve the clarity of causal analysis.

- **Graph Embedding:** Applying graph embeddings can offer richer insights into the graph's structure. They generate low-dimensional representations of nodes, preserving their structural relationships and enabling discovery of hidden relationships to form a complete KG.

- **Scalability Solutions:** Adopting more efficient algorithms and leveraging distributed computing can help manage the scalability issues associated with large datasets.

- **Automating context extraction:** This will help automating the process of identifying highly relevant information that could give causal insights behind an input data.

In order to improve the efficiency of analysis, the next research question (RQ2) was formed. The next RQ explores a different approach to overcome the challenges and provide more robust insights into the causal relationships within the dataset.

Chapter 5

RQ2: Can we infer causality by extracting context from KGs represented with entities as nodes together with sentiment information?

Following the limitations identified in RQ1 regarding the complexity and scalability of KGs for causal analysis, the second research question (RQ2) investigates whether generating embeddings and incorporating Graph Neural Networks (GNNs) can provide more robust and efficient insights. This approach aims to improve the scalability, accuracy, and contextual understanding of causal relationships within the dataset. This aims to extract the relevant contextual information from the dataset through KG to analyse causes based on Cosine Similarity.

Tailoring the Solution to Address RQ2

To address RQ2, the proposed solution involves embedding the graph triplets into a 2 dimensional vector space and enhancing it further with GNNs, specifically utilizing the Deep Recursive Graph Infomax (DRGI) model. Additionally, the construction approach for the KG has been refined to enhance the representation of entities and relationships, and sentiments are incorporated differently to provide a more nuanced analysis. Two variations of solution to address RQ2 are explored,

Variation 1: Standard Entity Embedding and GNN Integration

This approach involves integrating standard entity embeddings with a Graph Neural Network (GNN). The entities extracted from the tweets are embedded using pre-trained Word2Vec models, creating a dense vector representation for each entity. These embeddings capture semantic similarities between entities based on their co-occurrences in large text corpora. A graph is then constructed where nodes represent entities and edges are

formed based on the cosine similarity between their embeddings, weighted by sentiment values derived from the tweets. The GNN is trained on this graph to enhance the node embeddings by learning from the graph structure, effectively capturing both the relational and sentiment information embedded in the tweets. Sentiment is integrated into the graph by assigning a sentiment score to each entity and calculating the differences in sentiment between connected entities. This method aims to reveal the causal relationships and influence patterns within the dataset by examining the enhanced node embeddings and their interactions.

Variation 2: Enhanced Sentiment Integration with Entity Embedding and GNN

Building upon the first variation, this method further incorporates sentiment words and their impacts into the graph construction and GNN training process. Alongside entities, sentiment-causing words are identified and embedded using Word2Vec, enriching the semantic representation. The graph is constructed to include additional edges that capture the relationships between entities and sentiment words, offering a more detailed view of sentiment interactions. The GNN model is then trained on this enriched graph, taking into account both entity relationships and the influence of sentiment words. This enhanced integration captures more detailed sentiment interactions by explicitly representing sentiment words and their influence on entities, aiming to provide a deeper understanding of how sentiments propagate and interact within the tweet network. This offers more nuanced insights into the causal dynamics present in the dataset.

These variations provide insights into the efficiency and utility of the models for causal analysis.

5.1 Solution Statement

The proposed solutions for RQ2 involves the following steps:

Version 1:

- **Data Preprocessing**: Clean and prepare the dataset for analysis.
- **Entity Extraction**: Identify and extract entities from the dataset.

- **Sentiment Mapping**: Map sentiment values to the extracted entities.
- **KG Construction**: Construct the KG with entities as nodes and relationships as edges.
- **Embedding Generation with Word2Vec**: Generate entity embeddings using the Word2Vec model.
- **Graph Neural Network (GNN) Integration**: Integrate GNNs, particularly the DRGI model, to enhance the embeddings.
- **Causal Tweet Identification**: Identify and analyze causal relationships using the enhanced embeddings.

Version 2:

- **Data Preprocessing**: Clean and prepare the dataset for analysis.
- **Entity Extraction**: Identify and extract entities from the dataset.
- **Sentiment Mapping**: Map sentiment values to the extracted entities.
- **KG Construction**: Construct the KG with entities as nodes and relationships as edges.
- **Embedding Generation with Word2Vec**: Generate entity embeddings using the Word2Vec model.
- **Graph Neural Network (GNN) Integration**: Integrate GNNs, particularly the DRGI model, to enhance the embeddings by incorporating sentiment words and their impacts.
  - **Sentiment-Causing Words**: Identify sentiment-causing words along with entities, and embed them using Word2Vec to enrich the semantic representation.

  - **Graph Construction**: Include additional edges in the KG to capture relationships between entities and sentiment-causing words, offering a more detailed view of sentiment interactions.
  - **GNN Training**: Train the DRGI model on this enriched graph, taking into account both entity relationships and the influence of sentiment-causing words.

- **Causal Tweet Identification**: Identify and analyze causal relationships using the enhanced embeddings.

5.2 Dataset

The Annotated Corpus for Named Entity Recognition dataset used for the first research question (RQ1) is also utilized for the second research question (RQ2). The second research question aims to determine whether causal insights can be gathered more effectively by constructing KGs based on entities from tweets as nodes, linked through contextual similarities, with sentiments expressed as edge weights. This dataset [57] is highly relevant and effective for addressing RQ2 due to its detailed annotations and rich contextual information.

Relevance to RQ2

The Annotated Corpus for Named Entity Recognition dataset is particularly well-suited for this research question for several reasons:

Detailed Entity Information

The dataset provides comprehensive annotations, including Part-of-Speech (POS) tags and Named Entity Recognition (NER) tags for each word in the sentences. This detailed information allows for accurate identification and extraction of entities, which are essential for constructing a KG.

Rich Contextual Data

The sentences in the dataset provide rich contextual information that is crucial for understanding the relationships between entities. By preprocessing the text to remove

stopwords and perform lemmatization, the dataset becomes even more suitable for analysis, ensuring that the entities and their relationships are captured effectively.

Sentiment Association

The dataset includes sentiment annotations that can be mapped to numerical values, enabling the addition of sentiment information as edge weights in the KG. This aspect is critical for understanding the emotional tone of the relationships between entities and for performing a nuanced analysis of the data.

Effective Use for Causal Analysis

The comprehensive structure and detailed annotations of the dataset facilitated an effective causal analysis:

- **Contextual Similarities**: By using cosine similarity to form relationships between entities, the constructed KG captured contextual similarities effectively. This approach ensured that entities with similar contexts were connected, providing a robust framework for causal analysis.

- **Sentiment-Enhanced Relationships**: The incorporation of sentiment information as edge weights allowed for a more nuanced understanding of the relationships. By analyzing the sentiment weights, deeper insights into the emotional dynamics between entities were obtained.

- **Advanced GNN Models**: The use of advanced GNN models, such as DRGI, enabled the extraction of complex and subtle causal relationships. These models leveraged the rich contextual and sentiment information provided by the dataset, resulting in a comprehensive and accurate causal analysis.

In summary, the Annotated Corpus for Named Entity Recognition dataset is highly suitable for addressing RQ2 due to its detailed entity annotations, rich contextual information, and sentiment mappings. These features facilitated the construction of a sentiment-enhanced KG, enabling a thorough and nuanced causal analysis. The dataset's structure and annotations provided a solid foundation for employing advanced GNN models to extract meaningful insights, making it an invaluable resource for this research question.

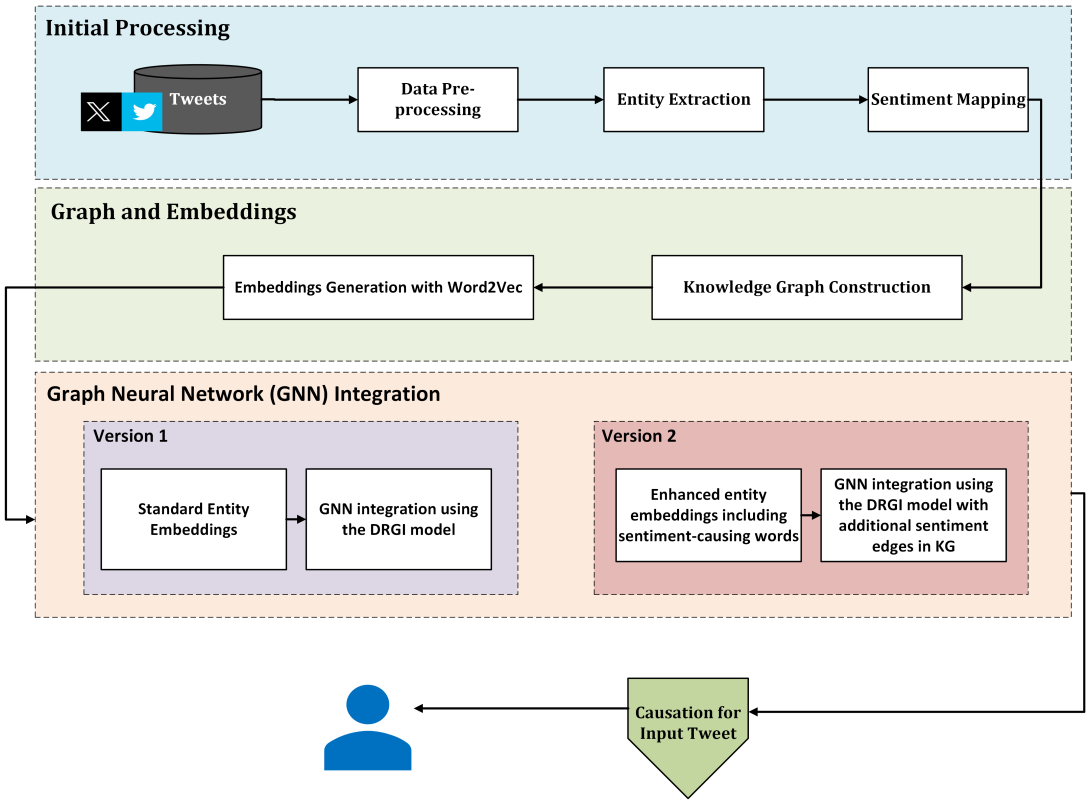

Figure 5.1: RQ2: Methodology for Application of GNN to KG Incorporated with Sentiments to infer Causality

5.3 Methodology

This section details the steps for implementation of the 2 versions as presented in the solution statement.

5.3.1 Data Preprocessing

The structured dataset containing sentences and sentiment tags, is preprocessed to convert the text into a structured format suitable for analysis. This involves tokenizing the text, removing stop words, and lemmatizing the tokens as detailed below including an example as they are performed.

- **Tokenization**: Each sentence in the dataset is split into individual words or tokens.
  - Example: The sentence "Thousands of demonstrators have marched through London" is tokenized into ["Thousands", "of", "demonstrators", "have", "marched", "through", "London"].

- **Stop Word Removal**: Common words that do not contribute to the meaning (e.g., "and," "the," "is") are removed as shown in the next step of processing the example.
  - Example: ["Thousands", "demonstrators", "marched", "London"].
- **Lemmatization**: Words are reduced to their base or root form (e.g., "marched" becomes "march") as shown below after processing further.
  - Example: ["thousand", "demonstrator", "march", "London"].

5.3.2 Entity Extraction

Entities are extracted from the preprocessed text, and their types are identified. This involves iterating through each token in the text and using named entity recognition techniques to categorize them.

- **Named Entity Recognition (NER)**: Identify entities such as people, organizations, locations, and other proper nouns.
  - Example: From the sentence "Thousands of demonstrators have marched through London," the entity "London" is identified as a location.
- **Categorization**: Entities are categorized into different types, such as geographical entities, organizations, and persons.
  - Example: The entity "London" is categorized as a geographical entity.

5.3.3 Sentiment Mapping

The sentiment of each entity is mapped using a predefined sentiment mapping scheme. This step assigns numerical sentiment values to each entity based on the sentiment of the sentences they appear in.

- **Sentiment Analysis**: Each sentence is analyzed to determine its sentiment, classified as positive, negative, or neutral.

  - Example: The sentence "Thousands of demonstrators have marched through London" may be classified as neutral.

- **Sentiment Mapping**: Entities within the sentences are assigned sentiment values based on the overall sentiment of the sentence.

  - Example: The entity "London" is assigned a neutral sentiment value (0).

- **Calculation**: For each entity, the average sentiment is calculated by summing the sentiment scores and dividing by the count of occurrences.

  - Example: If "London" appears in 10 sentences with varying sentiments, its average sentiment score is calculated.

5.3.4 Knowledge Graph Construction

A KG is constructed using NetworkX, where nodes represent entities and edges capture relationships based on similarity and sentiment.

- **Graph Initialization**: Create a directed graph where each node represents an entity.

  - Example: Initialize an empty graph object.

- **Edge Creation**: For each pair of entities, calculate the cosine similarity between their embeddings. If the similarity exceeds a threshold, an edge is added between the nodes.

  - Example: Calculate similarity between embeddings of "London" and "Iraq" and add an edge if the similarity is above the threshold.

- **Sentiment as Edge Weight**: The sentiment difference between entities is used as an additional attribute for the edges, capturing the sentiment dynamics between entities.

  - Example: If "London" has a sentiment score of 0.5 and "Iraq" has -0.5, the sentiment difference is used as an edge attribute.

5.3.5 Embedding Generation with Word2Vec

Generate entity embeddings using the Word2Vec model, which captures the contextual similarity between entities.

- **Training Word2Vec**: Use a large corpus of text to train the Word2Vec model, which learns vector representations of words.
  - Example: Train Word2Vec on a large text corpus to learn 300-dimensional embeddings for each word.
- **Embedding Extraction**: For each entity, retrieve its embedding from the trained Word2Vec model.
  - Example: Retrieve the 300-dimensional embedding for the entity "London."
- **Vector Representation**: These embeddings provide a numerical representation of the entities, capturing their semantic meaning.
  - Example: Each entity is now represented as a 300-dimensional vector.

5.3.6 Graph Neural Network (GNN) Integration

Version 1: Standard Entity Embedding and GNN Integration

Integrate GNNs, specifically the DRGI model, to enhance the embeddings and capture deeper structural relationships within the KG.

- **DRGI Model Architecture**: The DRGI model consists of multiple layers:
  - **Input Layer**: Accepts the entity embeddings.
  - **Graph Convolution Layers (GCNConv)**: Apply convolution operations on the graph to aggregate information from neighboring nodes. The formula for a single graph convolution layer is:

$$h_i^{(k+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{\sqrt{d_i d_j}} W^{(k)} h_j^{(k)}\right)$$

where $h_i^{(k+1)}$ is the representation of node i at layer $k+1$, $\mathcal{N}(i)$ is the set of neighbors of node i, d_i and d_j are the degrees of nodes i and j, $W^{(k)}$ is the weight matrix at layer k, and σ is an activation function (e.g., ReLU).

  - **Output Layer**: Produces enhanced entity embeddings.
  - Example: The DRGI model is structured to perform convolution operations on the KG and generate enhanced embeddings.

- **Graph Conversion**: Convert the KG to a format suitable for input to the GNN, such as PyTorch Geometric's Data object, which includes node features, edge indices, and labels.
  - Example: Convert the graph to a PyTorch Geometric Data object with node features representing the embeddings and edge indices representing the graph structure.

- **Training**: Train the DRGI model using a loss function (e.g., Mean Squared Error) to minimize the difference between the predicted and actual node embeddings.
  - Example: Train the model for a specified number of epochs, updating the model weights to minimize reconstruction loss.

Version 2: Enhanced Sentiment Integration with Entity Embedding and GNN

Building upon Version 1, this variation further incorporates sentiment words and their impacts into the graph construction and GNN training process.

- **Sentiment-Causing Words**: Identify sentiment-causing words along with entities, and embed them using Word2Vec to enrich the semantic representation.
  - Example: Alongside entities like "London," sentiment-causing words such as "happy" or "angry" are embedded and included in the KG.

- **Graph Construction**: Include additional edges in the KG to capture relationships between entities and sentiment-causing words, offering a more detailed view of sentiment interactions.

  - Example: Edges are added between "London" and sentiment-causing words based on their contextual relationships.

- **GNN Training**: Train the DRGI model on this enriched graph, taking into account both entity relationships and the influence of sentiment-causing words.

  - Example: The GNN model is trained to capture the impact of sentiment-causing words along with standard entity embeddings, leading to a richer semantic representation and better understanding of sentiment interactions.

5.3.7 Causal Tweet Identification

Identify and analyze causal relationships using the enhanced embeddings produced by the GNN.

- **Similarity Calculation**: Calculate the cosine similarity between the enhanced embeddings of tweets to identify similar tweets.

  - Example: Compute cosine similarity between the embedding of a given tweet and all other tweets in the dataset.

- **Top Similar Tweets**: For a given tweet, identify the top similar tweets based on the highest similarity scores.

  - Example: Identify the top 5 most similar tweets based on cosine similarity scores.

- **Causal Analysis**: Analyze the context and sentiment of the similar tweets to identify potential causal relationships.

  - Example: Examine the content and sentiment of the identified tweets to infer causal relationships between events.

5.4 Results and Analysis

The results presented here are based on two different versions of the methodology to provide a detailed comparative analysis. These versions demonstrate the methodology and the types of insights that can be derived from the data, that may help doing causal analysis.

5.4.1 Version 1 of Solution

In Version 1, the process was aimed at extracting causal relationships between tweets based on entity embeddings and sentiment scores. Below are the detailed results for a sample input tweet "Bedfordshire police said Tuesday that Omar Khayam was arrested in Bedford for breaching the conditions of his parole.":

- **Entity Embeddings and Sentiment Scores:** The method began by embedding entities extracted from tweets and mapping them with sentiment scores to capture both semantic and emotional contexts. This combined approach aimed to provide a richer representation of the data.

- **Cosine Similarity Calculation:** For the input tweet, "Bedfordshire police said Tuesday that Omar Khayam was arrested in Bedford for breaching the conditions of his parole," cosine similarity scores were calculated between this tweet and other tweets in the dataset. The goal was to identify tweets with similar semantic content and sentiment context.

- **Ranking and Selection:** Based on the cosine similarity scores, tweets that were most similar to the input tweet were ranked. The top five tweets with the highest similarity scores were selected as potential causes.

- **Results:**

  - **Cause: 0.9574**

    Tweet: The provincial governor must still sign the bill before it becomes law, a step seen only as a formality.

- **Cause: 0.9607**

  Tweet: The opposition has denounced the measure, comparing it to the draconian rule of the former Taleban in neighboring Afghanistan.

- **Cause: 0.9633**

  Tweet: Hardline lawmakers in Pakistan's North West Frontier Province have pushed through a law that aims to ensure "Islamic correctness" in public places and establishes a morality police to enforce decent behavior.

- **Cause: 0.9857**

  Tweet: In Saturday's elections, voters will cast ballots in nine provinces where no candidate won a majority in the previous round of voting.

The results in Version 1 show that the methodology effectively identifies tweets that are contextually and semantically related to the input tweet. The high similarity scores suggest strong relevance, indicating potential causal relationships. For instance, the tweet about the provincial governor signing a bill highlights a formal step in law enforcement, which could relate to the legal conditions mentioned in the input tweet. The tweet about the Taleban's rule in Afghanistan shows a comparison to stringent rules, potentially paralleling the conditions of parole mentioned.

These top results provide a deeper insight into the factors influencing the sentiment and context of the input tweet. By examining the top five tweets, one can gather comprehensive information about the causality behind the input tweet's sentiment. This combined analysis can help identify patterns and underlying themes that are crucial for understanding the causal dynamics within the dataset.

5.4.2 Version 2 of Solution

Version 2 introduced a comparative analysis using two different similarity measurements: cosine similarity based on TF-IDF vectors and sentiment-based similarity using enhanced node embeddings. The detailed results for the same input tweet are as follows:

- **Input Tweet:** Bedfordshire police said Tuesday that Omar Khayam was arrested in Bedford for breaching the conditions of his parole.

- **Most Similar Tweet based on Cosine Similarity:**

  Similarity: 0.1941

  Tweet: Officials say she died Tuesday.

- **Most Similar Tweet based on Sentiments:**

  Similarity: 0.9667

  Tweet: Representatives from the Asia Pacific Economic Cooperation Business Advisory Council are holding meetings this week to finalize their annual report for APEC leaders who will hold a summit on September 8 and 9.

These results show a significant difference between the two similarity measures. While the cosine similarity based on TF-IDF vectors identifies a tweet with low contextual relevance, the sentiment-based similarity finds a tweet with a higher relevance in terms of the sentiment and context. The tweet about APEC meetings is contextually rich and shares a thematic relevance with the input tweet, emphasizing the importance of incorporating sentiment information for more accurate causal analysis.

These results provide insights into the effectiveness of the proposed methodologies for causal analysis. The first version demonstrated the capability to extract meaningful causal relationships using entity embeddings and sentiment scores. The second version highlighted the importance of comparing different similarity measures to improve the accuracy and relevance of the extracted relationships.

These findings underline that combining various methodologies, such as embedding-based similarity with sentiment analysis and traditional cosine similarity, can enhance the understanding of causal relationships within the dataset. This approach ensures a more comprehensive analysis, yielding the right set of context to focus on, thus providing a robust framework for causal analysis in complex datasets.

5.5 Limitations and Transitioning to RQ3

While the methodologies and variations introduced in the implementation of the solutions for RQ2 represent significant advancements over RQ1, there are still several shortcomings

that need to be addressed to further enhance the efficiency and accuracy of causal analysis from tweets.

- **Entity Identification Improvement:** One of the critical steps in constructing an effective KG is the accurate entity identification from tweets. The current approach relies heavily on pre-trained Word2Vec models for entity embeddings and basic NLP techniques for entity extraction. However, this method can miss nuanced entities or incorrectly categorize them due to the informal and often ambiguous nature of social media text. Incorporating more sophisticated entity recognition algorithms and leveraging advanced NLP models such as BERT or GPT-3 can improve the accuracy and comprehensiveness of entity identification. These models can better capture the context and semantics of the text, leading to more precise and complete entity extraction.

- **Context Identification for Causality Extraction:** Identifying the correct context around entities is crucial for understanding causality. While RQ2 introduced sentiment-causing words to enrich the graph, capturing the broader context in which entities appear remains challenging. Contextual windows around entities, which include surrounding words and sentences, are essential for extracting causality. Future work should focus on developing methods to dynamically adjust these context windows and ensure they capture relevant information. This can involve advanced techniques such as contextual embeddings or attention mechanisms that highlight significant parts of the text related to the entities.

- **Improved KG Construction:** Although RQ2 provided a more nuanced view by integrating sentiment-causing words, the overall construction of the KG still has room for improvement. Ensuring that the KG accurately represents the relationships between entities and their contexts is vital. Enhancing the KG construction process by incorporating richer, more detailed entity relationships and dynamically updating the graph as new data comes in is essential. This will ensure the KG remains current and reflective of the latest trends and interactions on social media.

- **Introduction of LLMs for Enhanced Entity Relationship Identification:** Large Language Models (LLMs), such as GPT-3, can significantly enhance entity relationship identification within the KG. These models can understand and generate human-like text, making them ideal for capturing complex relationships and interactions between entities. By integrating LLMs into the KG construction process, we can achieve a more accurate and comprehensive representation of the data, which is crucial for effective causal analysis.

- **Context Extraction through Retrieval-Augmented Generation (RAG):** RAG models combine retrieval-based and generative approaches to provide contextual information. This method can improve context extraction by retrieving relevant documents or text fragments and using them to generate responses that consider a broader context. Incorporating RAG models can enhance our ability to identify causal relationships by providing more accurate and contextually rich information.

- **Dynamic KG Updating with Neo4j DB:** For social media data, where new information is continuously generated, dynamically updating the KG is vital. Using a graph database like Neo4j allows for efficient real-time updates and queries. This capability ensures that the KG remains up-to-date with the latest data, providing a more accurate basis for causal analysis. Neo4j's capabilities in handling complex and large-scale graph data make it an ideal choice for this purpose.

Although there has been progress from RQ1 to RQ2, where the focus shifted from providing various representations to analyzing causes, the approach was refined by extracting context windows to search for causes. Future efforts will aim to make this process more efficient by improving the construction of knowledge graphs (KG), which serves as the foundation for extracting causes. This improvement is the central focus of RQ3, where large language models (LLMs) are introduced to enhance entity-relationship identification for more effective KG construction. This not only improves the relevance of the extracted context but also enhances the accuracy of reasoning in responses with the help of a generative LLM. Additionally, RQ3 addresses the need for dynamic updates to the KG, a

critical factor for social media analysis, by incorporating a Neo4j database to ensure the KG stays current with evolving information.

Chapter 6

RQ3: How can the construction and updating of KGs be enhanced using LLMs and dynamic graph databases to improve causal analysis of social media data?

Following the advancements and limitations identified in RQ2, RQ3 focuses on leveraging the capabilities of LLMs and dynamic graph databases to enhance the construction and updating of KGs. This approach aims to improve entity relationship identification, context extraction, and dynamic updating to provide a more accurate and scalable solution for causal analysis.

This research question has been thoroughly investigated, and the solution is proposed through a model called **PRAGyan**. A paper detailing this model, titled **"PRAGyan - Connecting the Dots in Tweets"**, has been accepted at the 16th International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2024). This study compares the proposed model with a baseline LLM for context retrieval, which will be discussed in detail in this section.

6.1 Solution Statement

To address RQ3, we developed a method that combines KGs with LLMs to analyze the causes behind various events and statements on social media platforms. Our approach specifically focuses on utilizing the rich semantic relationships and temporal information encoded in KGs to enhance the depth of causal analysis. We used a RAG model, with the KG stored in a Neo4j format, referred to as PRAGyan, to retrieve relevant context for causal reasoning. By comparing the results from our KG-enhanced LLM RAG model to a baseline LLM (GPT-3.5 Turbo), we observed significant improvements as the data corpus expanded.

Through our qualitative analysis, we identified that the integration of KGs and LLMs provides better interpretability and actionable insights, which are critical for informed decision-making across various domains. Additionally, our quantitative analysis using BLEU and cosine similarity metrics showed that our approach outperforms the baseline model by 10%. This demonstrates the effectiveness of our method in delivering a deeper understanding of causal relationships from social media data.

6.2 Dataset

For the third research question (RQ3), which explores the integration of KGs and LLMs to enhance causal reasoning capabilities on social media data, the COVID-19 Tweets dataset was employed. This dataset (sample shown in Table 6.1) from [17] provides a rich source of real-time updates, personal experiences, and opinions shared by individuals on Twitter, making it highly suitable for the intended analysis.

Table 6.1: Sample Subset of Tweets on Covid19

Date	**Text**
2020-07-25 12:27:21	If I smelled the scent of hand sanitizers today on someone in the past, I would think they were so intoxicated that the alcohol was oozing out of their pores YUCK #COVID19
2020-07-25 12:27:17	Hey @Yankees @YankeesPR and @MLB - wouldn't it have made more sense to have the players pay their respects to the Americans who died from #COVID19 rather than show support for a #Marxist organization like #BlackLivesMatter? Our nat'l crisis isn't about race. It's the #pandemic.
2020-07-25 12:27:14	Trump never once claimed #COVID19 was a hoax. We all claim that this effort to blame Trump for it instead of the real instigator, the Communist China government, has always been & remains the hoax. He acted early to stop travel from China, then Europe. Continues to help states.
2020-07-25 12:27:10	The one gift #COVID19 has give me is an appreciation for the simple things that were always around me ... my yard being one of them
2020-07-25 12:27:06	#coronavirus #covid19 deaths continue to rise. It's almost as bad as it ever was. Politicians and businesses want your money. But it is not safe. #STAY-atHOME
2020-07-25 12:27:00	You now have to wear face coverings when out shopping - this includes a visit to your local Community Pharmacy #Covid19 #WearAMaskPlease

COVID-19 Tweets Dataset Overview

Twitter, now known as X, is a primary platform for individuals to share real-time updates, personal experiences, and opinions on various topics. The COVID-19 Tweets dataset was obtained from Kaggle [17] and provides a comprehensive collection of tweets related to the COVID-19 pandemic. This dataset has been widely used for data cleaning and learning purposes and has more than 22,000 downloads, indicating its popularity and relevance for

research.

Dataset Features:

The dataset contains thirteen features/fields, of which only a subset was used for this study. The relevant fields include:

- **user_name**: The username of the individual who posted the tweet.
- **user_location**: The location of the user, which can provide geographical context.
- **user_description**: A brief description provided by the user.
- **user_created**: The date when the user's account was created.
- **user_followers**: The number of followers the user has.
- **user_friends**: The number of friends the user has.
- **user_favourites**: The number of tweets the user has favorited.
- **user_verified**: A boolean indicating whether the user's account is verified.
- **date**: The timestamp when the tweet was posted.
- **text**: The content of the tweet.
- **hashtags**: The hashtags used in the tweet.
- **source**: The source from which the tweet was posted (e.g., mobile, web).
- **is_retweet**: A boolean indicating whether the tweet is a retweet.

For this study, the focus was primarily on the **text** and **date** fields. The **text** field contains the tweet content, which is crucial for sentiment analysis and understanding the context of the tweets. The **date** field provides the timestamps, allowing for the analysis of temporal patterns and trends in the data.

Dataset Collection and Coverage

The tweets were gathered between February 28, 2020, and July 23, 2020, using the Twitter API with the hashtag #covid19. This period covers significant events and developments during the early months of the COVID-19 pandemic, making the dataset a valuable resource for analyzing public sentiment and reactions over time. The dataset has a global coverage, capturing tweets from various parts of the world, which provides a diverse and comprehensive view of the pandemic's impact on different regions.

Dataset Statistics

The dataset contains over 35,000 rows, with an average tweet length of approximately 17 words after preprocessing. The total size of the dataset is 68.71MB. The preprocessing steps involved tokenizing the text, removing stopwords, and performing lemmatization to ensure that the data is clean and ready for analysis.

Relevance to RQ3

The COVID-19 Tweets dataset is particularly relevant for RQ3 due to several reasons:

Real-Time and Contextual Data

The dataset provides real-time updates and personal experiences related to the COVID-19 pandemic, offering rich contextual information. This real-time data is crucial for understanding the dynamic nature of public sentiment and reactions, which is essential for performing causal reasoning.

Temporal Coverage

The dataset spans several months of the early pandemic period, allowing for the analysis of temporal patterns and trends. This temporal aspect is vital for understanding how sentiments and opinions evolved over time and for identifying potential causal relationships between events and public reactions.

Rich Sentiment Data

The tweets contain a wide range of sentiments expressed by users, from fear and anxiety to hope and resilience. This diversity in sentiment data provides a robust foundation for sentiment analysis and for understanding the emotional tone behind the tweets, which is critical for causal reasoning.

Diverse User Base

The dataset captures tweets from users across the globe, offering a diverse and comprehensive view of the pandemic's impact. This geographical diversity allows for the testing of regional differences in public sentiment and reactions, adding another layer of depth to the analysis.

Integration with KGs and LLMs

The dataset's rich textual content and temporal coverage make it an ideal candidate for integration with KGs and LLMs. By constructing a KG from the tweets and using LLMs to analyze the data, the research can leverage the strengths of both approaches to enhance causal reasoning capabilities. The KG can capture the relationships between entities and events, while the LLMs can provide contextual understanding and generate detailed explanations.

In summary, the COVID-19 Tweets dataset is highly suitable for addressing RQ3. Its rich contextual data, temporal coverage, diverse user base, and detailed sentiment information provide a robust foundation for integrating KGs and LLMs to enhance causal reasoning capabilities. This dataset allows for a comprehensive analysis of public sentiment and reactions during the early months of the COVID-19 pandemic, providing valuable insights for understanding the factors driving social media interactions.

6.3 Methodology

Our study design consists of two main approaches: the baseline method and the proposed model. Each approach addresses the issue of causal reasoning in social media data,

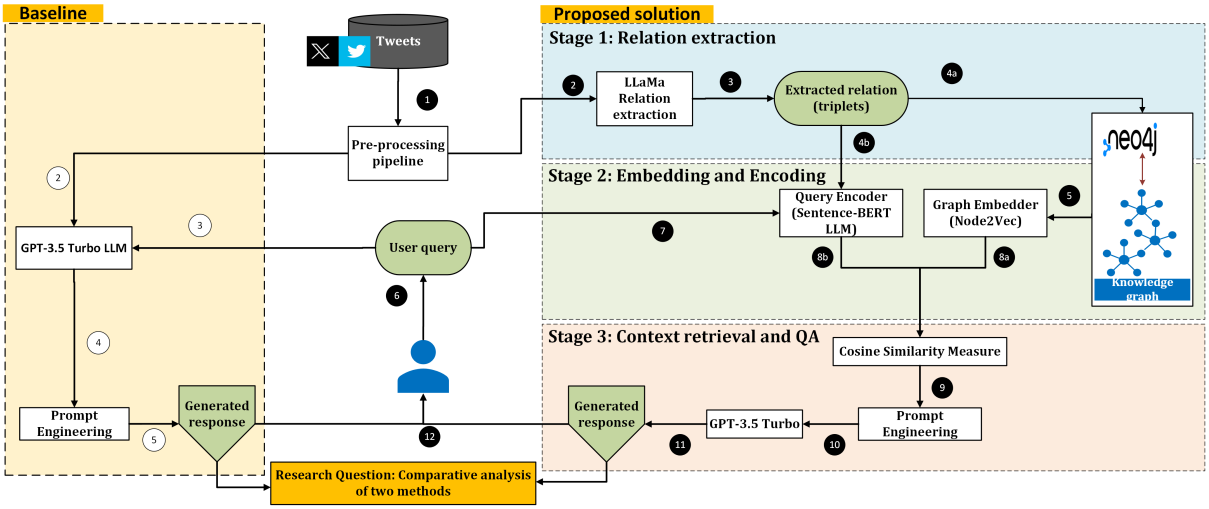

Figure 6.1: RQ3: Methodology for Baseline and Proposed Solution for Causal Analytics through RAG

particularly tweets, using different methodologies.

The technologies utilized in these methods were carefully selected for their efficiency and ease of accessibility in handling the specific tasks. The proposed model is an experimental architecture that integrates several technologies in a coordinated flow to optimize performance. This approach aims to enhance the accuracy of results by leveraging the strengths of each technology, combining them to create an effective solution.

6.3.1 Pre-processing

The initial step (**Step 1**) is data pre-processing, which is common to both the baseline method and our proposed method. Besides the text of the tweets, we also extract timestamps and ensure their accuracy by handling various time zones, correcting any inconsistencies, and standardizing the format. Tweets are typically unstructured, ambiguous, and noisy (with special characters, acronyms, and URLs). Therefore, we first use a dictionary to expand contractions commonly used in social media slang, replacing them with their full forms. We also remove special characters, emojis, URLs, and hashtags.

The Baseline:

The baseline model employs the GPT-3.5 Turbo LLM to directly query the entire preprocessed text corpus through prompt engineering for causal reasoning when a user query is provided. Unlike the proposed model, this method does not utilize a KG for

context retrieval but instead relies solely on the LLM's capability to process and generate responses based on the raw text data.

Proposed Model (PRAGyan):

The proposed PRAGyan model follows a three-stage process: Relation extraction, Embedding/Encoding, and Context Retrieval and QA. Each stage utilizes specific technologies chosen for their ability to enhance the overall process.

6.3.2 Relation Extraction

In the proposed model, relation extraction is carried out using the LLaMa3 model (**Step 2**), specifically the fine-tuned variant from Hugging Face (solanaO/llama3-8b-sft-qlora-re). This model identifies entities and their relations (triplets) from the pre-processed tweets (**Step 3**). The process starts by loading the pre-trained model with a PEFT adapter for efficient memory usage. The tokenizer creates prompts from the preprocessed text, and the model generates outputs, forming the triples. These are stored as a KG within a graph database, Neo4j (**Step 4a**). Additionally, Sentence-BERT encodings are obtained for all the relational triples and stored as edge property information in the database (**Step 4b**).

6.3.3 Embedding and Encoding

In this stage, the KG stored in Neo4j is embedded using Node2Vec (**Step 5**). This method captures both local and global structural information of the graph, creating high-quality vector representations stored in Neo4j for efficient retrieval of contextually relevant data.

The objective is to maximize the probability of observing a node's neighbors given its embedding through the Node2Vec algorithm:

$$\max \sum_{u \in V} \sum_{v \in N(u)} \log \Pr(v|u)$$

where $\Pr(v|u)$ is the probability of node v being a neighbor of node u, computed using the softmax function:

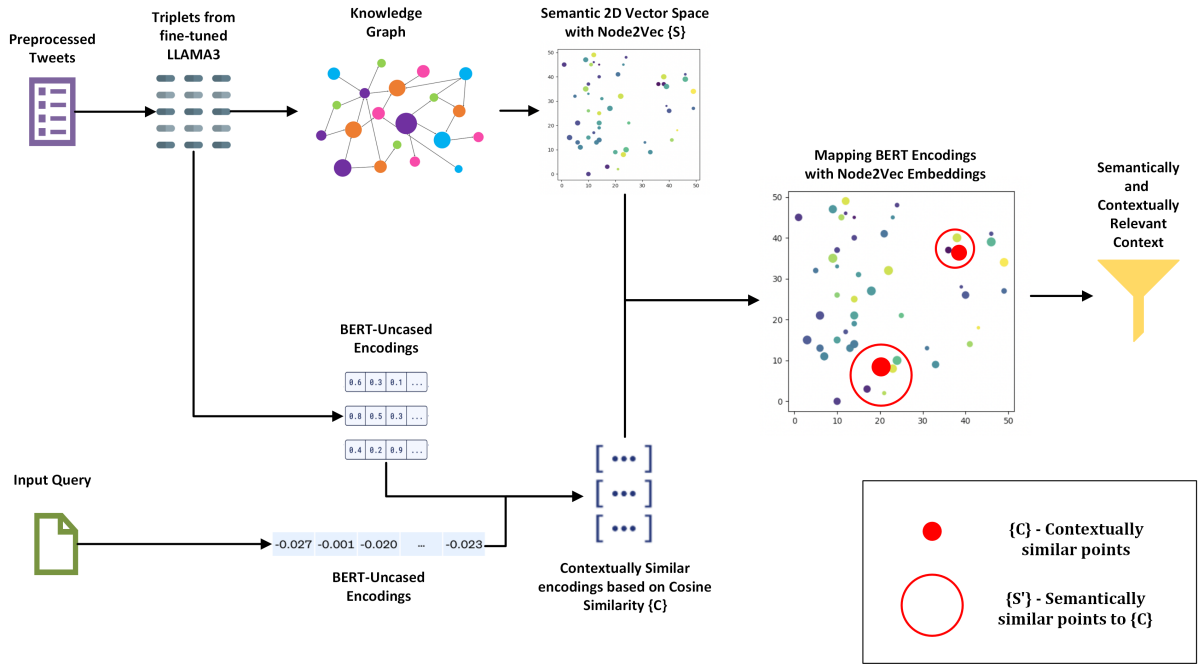

Figure 6.2: RQ3: Generation of Embeddings for KG and Encodings Input Query and Triples to perform Contextual and Semantic Similarity Calculations

$$\Pr(v|u) = \frac{\exp(\mathbf{z}_v^\top \mathbf{z}_u)}{\sum_{w \in V} \exp(\mathbf{z}_w^\top \mathbf{z}_u)}$$

Here, $\mathbf{z}_u$ and $\mathbf{z}_v$ are the embeddings of nodes u and v, respectively. The learned embeddings are stored in the graph database for efficient retrieval and analysis.

When a user inputs a query (**Step 6**) to reason causation, the query is encoded using an LLM (Sentence-BERT) to create a vector representation (**Step 7**). This representation is mapped onto the corresponding Node2Vec embeddings to identify the most relevant context in the next stage.

6.3.4 Context Retrieval and QA

This step involves retrieving the most relevant contexts from the KG to engineer prompts, a process known as prompt engineering (**Step 9**). These prompts are fed into a GPT-3.5 Turbo model (**Step 10**) to generate a response. The response, enriched with relevant context from the KG, is provided to the user, ensuring a detailed and comprehensive answer. This is known as RAG. It involves:

- *Identifying Contextually Relative Subgraph:* The Sentence-BERT encoding of the query is compared with the encodings calculated from the triples to choose a highly similar set through Cosine similarity. This set is then looked up on the Node2Vec embedding space with their corresponding embeddings. The resulting set of node embeddings is contextually similar to the query, which is passed on to the next step for further refinement to extract contextual information.

- *Retrieving Final Context - Contextually and Semantically Relevant:* With this set, using Cosine similarity, we identify its top semantically similar subsets of neighboring nodes and edges in the Node2Vec embedding space. Both similarity measures are validated for a threshold of 0.35, determined through trials and inspections of contexts with various thresholds. This ensures the right amount of context that could potentially provide the true cause. The subgraphs are mapped onto the pre-processed sentences, retrieving the top-ranked sentences (contextually and semantically most similar to the input statement) that serve as the source context for prompting with LLM. Including temporal constraints ensures the retrieved context respects the chronological order of events.

The retrieved context is then used by the GPT-3.5 Turbo LLM to generate responses with an effective prompt instructing the model to infer the cause for the input query.

6.4 Results and Analysis

The performance of the integrated approach, PRAGyan, was evaluated and compared against a baseline model, GPT-3.5 Turbo. This evaluation involved the use of manually designed queries derived from tweets, as well as accurately constructed causal relationships for validation. This ensured a comprehensive assessment of each model's capabilities. In total, 50 queries were created from the last 10,000 rows of the dataset. These queries were designed to analyze large portions of the tweets, requiring responses that drew on information spread across multiple tweets, rather than from a single tweet. This approach ensured that the models had to synthesize information from a collective set of tweets. The

ground truth for each query was established through this process, serving as a benchmark for comparing the models' responses, with these benchmarks considered to be the accurate answers against which the models were validated.

6.4.1 Qualitative Evaluation

Table 6.2: Responses from PRAGyan (Context from Embeddings through Selective Tweets) and the Baseline (Entire Corpus as the Text Source)

Question	Summary from Baseline	Summary with Context (PRAGyan)	Tweets
What measures are being taken to support patients recovering from COVID-19?	Measures are being taken to support patients recovering from COVID-19 due to the impacts of the pandemic and the need to address the health and well-being of individuals affected by the virus.	Efforts to enhance treatment, such as the setup of a plasma bank, prayer support, and sharing of updates on ICU patients indicate a focus on improving recovery outcomes for COVID-19 patients.	[2020-07-25] [12:21 PM] Rajasthan government today started plasma bank at Sawai Man Singh hospital in Jaipur for treatment of COVID-19. [2020-07-25] [12:25 PM] Mom is in ICU due to COVID just want prayer from you and everyone who is listening you COVID19. [2020-07-25] [12:27 PM] Praying for good health and recovery from COVID19.
How did the COVID-19 pandemic affect grocery shopping behaviors and store operations?	Fear of shortages and uncertainty due to the COVID-19 pandemic led to panic buying and stockpiling of essential items, such as food, toilet paper, and sanitizing products, in anticipation of potential disruptions in supply chains and daily routines.	The COVID-19 pandemic led to panic buying, resulting in empty shelves and increased demand for online shopping. Grocery stores implemented measures like dedicating shopping hours for the elderly and encouraging online orders. The pandemic caused shortages of essential items like toilet paper and hand sanitizer, and store employees faced increased pressure to keep shelves stocked.	[2020-03-15] [09:00 AM] Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak [2020-03-16] [09:25 AM] Due to the COVID-19 situation, we have increased demand for all food products... The wait time may be longer for all online orders, particularly beef share and freezer packs. [2020-03-16] [10:00 AM] As news of the region's first confirmed COVID-19 case came out of Sullivan County last week, people flocked to area stores to purchase cleaning supplies, hand sanitizer, food, toilet paper and other goods. [2020-03-16] [11:00 AM] All month there hasn't been crowding in the supermarkets or restaurants, however reducing all the hours and closing the malls means everyone is now using the same entrance and dependent on a single supermarket.

Qualitative evaluation involves using metrics such as faithfulness, answer relevance, and context relevance to gain a nuanced understanding of how effectively the models capture and convey meaningful information. Table 6.1 below explains how the baseline model processes the entire corpus without leveraging specific context from the KG, generating responses based solely on raw text data.

For the first question regarding measures to support COVID-19 patients, PRAGyan identifies specific actions such as setting up a plasma bank, offering prayer support, and providing updates on ICU patients. These detailed steps highlight concrete measures taken to enhance treatment outcomes, whereas the baseline model offers a more general-

ized summary lacking specific details. Similarly, for the question about the pandemic's impact on grocery shopping behaviors and store operations, PRAGyan details instances of panic buying, dedicated shopping hours for the elderly, and increased demand for online shopping, along with specific challenges faced by stores. In contrast, the baseline model provides a broad response, focusing on general fears of shortages and stockpiling without offering specific contextual examples. This comparison underscores PRAGyan's enhanced ability to deliver richer, context-aware responses by utilizing structured knowledge from the KG, thereby offering deeper insights and more actionable information.

6.4.2 Quantitative Evaluation

Table 6.3: Evaluation Metrics for the Baseline and PRAGyan Models

Metric	The Baseline	PRAGyan
BLEU	0.45	**0.49**
Cosine Sim with LLM Encoder	0.86	**0.88**

Quantitative evaluation is performed utilizing two metrics:

- **BiLingual Evaluation Understudy (BLEU) Score:** BLEU measures the precision of n-grams in generated text against reference texts, evaluating word or phrase matches with reference causes. It provides a quick, reproducible measure of text accuracy.

- **Cosine Similarity on Sentence-BERT Encodings:** The BERT Similarity metric captures contextual meanings and semantic similarity between texts, making it ideal for assessing the conveyed information rather than exact structure.

Together, these metrics ensure a comprehensive evaluation of both structural and semantic quality in generated text.

The two average metric scores for both models observed over multiple queries are presented in the table below. PRAGyan outperforms the baseline model across both

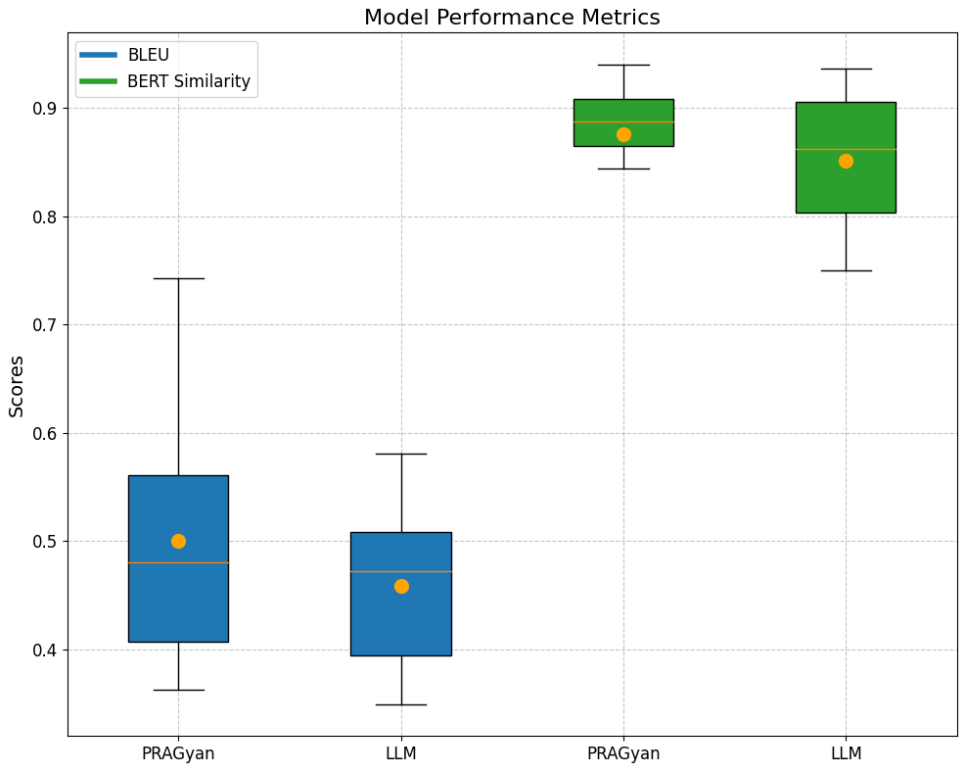

Figure 6.3: RQ3: Box plot showing metrics between the Baseline and PRAGyan

metrics, indicating its superior performance in generating responses that are both accurate and comprehensive.

Our proposed model leverages the strengths of both KG and RAG, resulting in better overall performance. It showed significant advantages for the COVID-19 tweets dataset, generating contextually and semantically relevant and accurate information, thus improving the quality of insights derived from this social media content. However, higher score variations suggest room for improvement in the KG's construction, as it currently only conveys correlation information.

In contrast, while the baseline provides more consistent performance, its generally lower scores indicate less accuracy. As the data environment grows, this increases the overhead for the baseline, potentially affecting results and processing times. Hence, maintaining a graph database helps with dynamic real-time social media updates and faster querying.

Hence, PRAGyan demonstrates superior average performance and the capability to produce highly relevant and accurate text, as indicated by higher metric scores. This

makes it particularly beneficial for analyzing social media data, where context and relevance are crucial. The variability in its performance highlights its potential to excel in many instances, suggesting that integrating LLM with KG through RAG enhances the model's ability to generate high-quality text.

6.4.3 Discussion

In addressing the limitations identified in the solutions for the previous two research questions, the proposed PRAGyan model offers a significant improvement in causal analysis from social media data, particularly tweets. This model enhances the scalability, accuracy, and contextual understanding of causal relationships within the dataset by integrating KGs and LLMs.

Some of the key improvements are:

- **Improved Entity and Relation Extraction**: By using the fine-tuned LLaMa3 model, PRAGyan effectively identifies and extracts entities and their relationships from tweets, ensuring a more accurate and comprehensive KG construction.

- **Enhanced Embedding and Encoding**: The use of Node2Vec for embedding the KG captures both local and global structural information, creating high-quality vector representations. This approach ensures efficient retrieval of contextually relevant data and keeps the embeddings up-to-date with the evolving graph.

- **Context Retrieval and QA**: The integration of context retrieval using Sentence-BERT and prompt engineering with GPT-3.5 Turbo ensures that the generated responses are enriched with relevant context from the KG. This method, known as RAG, significantly enhances the accuracy and relevance of the responses.

- **Dynamic and Real-Time Updates**: The use of Neo4j for storing and dynamically updating the KG ensures that the model can handle real-time updates and queries, making it well-suited for the ever-changing landscape of social media data.

The PRAGyan model's ability to combine these advanced techniques results in a robust framework that not only addresses the limitations of the previous solutions but also

provides a deeper understanding of causal relationships. This comprehensive approach offers a more scalable, accurate, and contextually aware analysis of social media data.

In conclusion, the PRAGyan model effectively answers the research question by demonstrating that integrating KGs with graph neural networks and advanced language models can significantly improve causal analysis in social media data. The model's superior performance in generating contextually rich and accurate responses underscores its potential as a powerful tool for understanding and analyzing the dynamics of social media interactions.

6.5 Limitations

The most efficient solutions were arrived at in the end with PRAGyan through various experiments and refinements of the RQs. Although it is a significantly improved solution to the primary problem statement, it has limitations.

- The queries used for evaluation did not span the entire dataset due to time and effort constraints, limiting the scope of the analysis.
- Ground truths were manually established without cross-validation by multiple peers, which introduces a potential risk of inaccuracies in the results, in terms of quantitative evaluation.
- Models like Node2Vec, Sentence-BERT, and GPT-3.5 Turbo can present challenges when working with different domains, differently formatted data, and varied working environments. They were chosen for their accessibility and compatibility with the existing resources and environment, making them practical for this research.
- Evaluating RAG and LLM models is complex due to their hybrid nature. RAG systems require an assessment of both retrieval and generation components, focusing on dynamic knowledge bases and subjective measures of relevance and coherence as explained in [30]. Similarly, [58] highlights that evaluating LLMs demands a comprehensive approach to address data leakage, inappropriate content, and safety standards.

- Metrics like BLEU and Cosine Similarity provide quantitative assessments but may not fully capture qualitative aspects of causal reasoning, necessitating human evaluation. Tailoring the framework with appropriate models and metrics ensures robust and contextually relevant evaluation across different datasets and use cases.

- Another limitation involves the dynamic nature of social media data. Continuous updates and the evolving linguistic landscape require the models and the KG to be frequently updated to maintain relevance and accuracy. This dynamic updating can be resource-intensive and may introduce latency issues in real-time applications.

- Lastly, the reliance on pre-trained models like GPT-3.5 Turbo, which are proprietary, can pose challenges related to transparency and reproducibility. Open-source alternatives and fine-tuning on specific datasets may offer better control but come with their own set of challenges related to computational resources and expertise.

In conclusion, while PRAGyan offers significant advancements in causal analysis of social media data, addressing its limitations through ongoing refinement, dynamic updating, and comprehensive evaluation approaches will be crucial for enhancing its effectiveness and applicability across various domains.

Chapter 7

Usecase

The structured dataset from Kaggle [59] is a comprehensive collection of Reddit comments related to the Israel-Palestine conflict. This dataset is updated daily, providing real-time insights into public sentiments and affiliations as expressed in online discussions. The dataset includes various fields that capture different aspects of the comments like timestamps, controversiality (score indicating how divisive or contentious a comment is) and subreddit (under which topic comments are posted), although only the timestamp column is utilized in this research, making it a valuable resource for analyzing social media discourse and understanding the dynamics of public opinion on this sensitive issue. By applying the solution developed for RQ3, which involves constructing and updating KGs using dynamic graph databases, we can efficiently manage and analyze this dataset. This approach allows for real-time tracking of sentiment shifts, identification of key influencers, and understanding the causal relationships between different events and public reactions.

For the specific use case of the Israel-Palestine conflict, causal analysis can reveal how specific events trigger changes in public sentiment and discourse. For instance, identifying which incidents lead to spikes in negative sentiment can help policymakers and conflict monitors understand the root causes of public outrage or support. Some more insights could include,

1. **Media Coverage of Violent Clashes**: Analyzing how the coverage of violent clashes, such as bombings or shootings, affects public sentiment can help understand which events are most likely to influence public opinion and how media portrayal impacts perceptions as described in [60].

2. **Political Statements and Policies**: Examining how statements from key polit-

ical leaders or the introduction of new policies (e.g., settlement expansions, peace negotiations) correlate with shifts in public sentiment can highlight the influence of political rhetoric and decisions on public attitudes [61].

3. **Social Media Trends**: Identifying specific hashtags, posts, or viral content that lead to peaks in social media activity and sentiment can provide insights into how digital discourse shapes and reflects public opinion during critical moments in the conflict [60].

4. **Humanitarian Crisis Reports**: Analyzing the impact of reports on humanitarian crises, such as blockades, medical shortages, or refugee situations, can help understand how information about human suffering influences public sentiment and advocacy efforts [62].

5. **Peace Talks and Ceasefires**: Evaluating how announcements of peace talks, ceasefires, or diplomatic efforts impact public sentiment can reveal the public's hope or skepticism towards conflict resolution efforts [63].

This deeper understanding can inform diplomatic strategies, media reporting, and conflict resolution efforts. In the modern world, such insights are invaluable for creating responsive and informed policies that address the underlying issues rather than just the symptoms. By uncovering the causal relationships within social media interactions, this analytical framework aims to serve as a powerful tool for navigating and mitigating the complex dynamics of international conflicts.

Dataset

The creation and maintenance of this dataset involve contributions from the author Asaniczka. The dataset covers comments from Reddit posts specifically related to the Israel-Palestine conflict, providing detailed temporal and geospatial coverage. A sample subset is shown in Table 7.1.

- **Temporal Coverage Start Date**: 10/06/2023

Comment Text	Created Time
A lot of leftists have been defending Hamas's actions in the past two days. You can see it all over Reddit.	2023-10-08 18:27:34
heals wayyy faster than a torn acl.. had kinda the same and was back ok the field 3 months - acl took me about 11 months to get back to 80-90%	2023-10-08 18:21:15
He's suffering. It's for his own good.	2023-10-08 18:00:44
Oh god, will they have to put him down?	2023-10-08 17:59:28
Not MCL, I tore my MCL and was back running in 3 months. ACL is completely different however.	2023-10-08 17:58:55
I literally just heard that the US is stationing amphibious assault ships as well as other marine corps assets.	2023-10-08 17:54:02
The first girl is German bruh	2023-10-08 17:45:18
Broke my leg that way last year, check my post history to see images	2023-10-08 17:41:53

Table 7.1: Sample Subset of Reddit Comments on the Israel-Palestine Conflict

- **Temporal Coverage End Date**: Ongoing (daily updates)
- **Geospatial Coverage**: Israel-Palestine

This temporal and geospatial coverage ensures that the dataset remains relevant and provides up-to-date information on the conflict. Understanding the provenance of the dataset is crucial for assessing its reliability and accuracy. The dataset was compiled by gathering comments using the Reddit API as mentioned in [59]. The comments were sourced from Reddit, focusing on fourteen subreddits that frequently discuss the current situation in Israel and Gaza. Posts related to the conflict were identified based on specific keywords, and then comments from those posts were extracted. This methodology ensures that the dataset is comprehensive and focused on discussions specifically related to the Israel-Palestine conflict. The dataset is released under the Open Data Commons Attribution License (ODC-By), which allows users to share and adapt the data, provided they give appropriate credit to the original creator. This license promotes open access to the dataset while ensuring that the contributions of the creator are acknowledged. It is updated daily, ensuring that it remains current and reflective of the latest discussions on Reddit. This frequent update schedule makes the dataset particularly valuable for real-time analysis and monitoring of public sentiment and discourse.

Description of the Fields

It includes several fields that capture different aspects of Reddit comments. Each field provides specific information that is crucial for a comprehensive understanding of the

data.

Table 7.2: Description of Fields from the Reddit Comments Dataset

Field Name	Type	Importance
comment_id	String	This field is essential for uniquely identifying each comment in the dataset, ensuring accurate data management and analysis.
score	Integer	The score indicates the popularity and reception of the comment within the Reddit community. It is a valuable metric for assessing the impact and visibility of a comment.
self_text	String	The text content is the primary source of information for analyzing the sentiments, opinions, and narratives expressed in the comments. It is crucial for text analysis and understanding the context of the discussions.
subreddit	String	This field helps categorize the comments based on the specific subreddits. It is useful for understanding the context and focus of the discussions within different communities on Reddit.
created_time	Datetime	The creation time is essential for temporal analysis, allowing researchers to examine trends and patterns over time. It helps in understanding the dynamics of the discussions and how they evolve.
post_id	String	This field links the comment to its parent post, providing context and enabling the analysis of comment threads. It is vital for understanding the relationship between posts and comments.
author_name	String	Analyzing the authors of comments provides insights into user behavior, engagement patterns, and community dynamics. This field is useful for studying the participation and influence of different users.
controversiality	Integer	The controversiality score indicates how divisive or contentious a comment is. It helps in identifying comments that generate strong reactions and debates within the community.
ups	Integer	This field provides a direct measure of the positive reception of a comment. It is useful for assessing the popularity and agreement with the comment.
downs	Integer	This field indicates the negative reception of a comment. It is valuable for understanding the dissent or disagreement with the comment within the community.

Application of PRAGyan for this Use Case

Applying the model developed for RQ3 to analyze this dataset provides significant insights into the public discourse surrounding the Israel-Palestine conflict. Only the created_time and self_text columns are considered from the dataset. For instance, when querying "What leads people to describe Gaza as being under siege similar to historical blockades?", the PRAGyan model provides the answer: "The query regarding people describing Gaza as being under siege similar to historical blockades is likely due to the ongoing conflict and humanitarian crisis in Gaza, characterized by extensive military operations and restrictions by various parties."

The query, "What leads people to describe Gaza as being under siege similar to historical blockades?", seeks to understand the factors contributing to this perception. To gather the context, selective comments that highlight specific incidents and sentiments were analyzed. The comments chosen by PRAGyan provide real-time, contextual relevance to

Table 7.3: Responses from PRAGyan (Context from Embeddings through Selective Tweets)

Question	Summary with Context (PRAGyan)	Reddit Comments
What leads people to describe Gaza as being under siege similar to historical blockades?	The query regarding people describing Gaza as being under siege similar to historical blockades is likely due to the ongoing conflict and humanitarian crisis in Gaza, characterized by extensive military operations and restrictions by various parties.	[2023-10-08] [18:27:34] A lot of leftists have been defending Hamas's actions in the past two days. You can see it all over Reddit. [2023-10-08] [18:18:35] Gaza is not an "open air prison" or a ghetto. It's under blockade by both Israel and Egypt, just like Japan was under US blockade during World War II. And it's under blockade due to hostile activities against Israel and Egypt, not just because like the Warsaw Ghetto. [2023-10-08] [14:34:31] Picture Gaza as a DMZ. [2023-10-08] [13:28:49] The last days of Gaza. Goodbye. [2023-10-06] [15:20:29] Free Palestine from the xenophobic and bigoted Israelis.

gather accurate insights:

- **Support for Hamas:** Comment from 2023-10-08 at 18:27:34: "A lot of leftists have been defending Hamas's actions in the past two days. You can see it all over Reddit." This tweet indicates public support and justification for Hamas's activities, contributing to the perception of Gaza being under siege.

- **Blockade Comparisons:** Comment from 2023-10-08 at 18:18:35: "Gaza is not an 'open air prison' or a ghetto. It's under blockade by both Israel and Egypt, just like Japan was under US blockade during World War II. And it's under blockade due to hostile activities against Israel and Egypt, not just because like the Warsaw Ghetto." This tweet draws parallels between Gaza's situation and historical blockades, emphasizing the severity of the blockade.

- **DMZ Comparison:** Comment from 2023-10-08 at 14:34:31: "Picture Gaza as a DMZ." This tweet likens Gaza to a demilitarized zone, further illustrating the perceived severity of the siege.

- **Dire Predictions:** Comment from 2023-10-08 at 13:28:49: "The last days of Gaza. Goodbye." This tweet underscores the extreme viewpoints driven by the humanitarian crisis in Gaza.

- **Anti-Israel Sentiment:** Comment from 2023-10-06 at 15:20:29: "Free Palestine from the xenophobic and bigoted Israelis." This tweet expresses strong anti-Israel sentiment, contributing to the perception of Gaza being under siege.

Summary with Context by PRAGyan

The context provided by the selected tweets helps in constructing a detailed summary. PRAGyan's summary states: "The query regarding people describing Gaza as being under siege similar to historical blockades is likely due to the ongoing conflict and humanitarian crisis in Gaza, characterized by extensive military operations and restrictions by various parties." The detailed inferences that can be made from this response are as follows,

- **Ongoing Conflict and Humanitarian Crisis:** The situation in Gaza involves extensive military operations and severe restrictions, which are central to the description of Gaza being under siege.
- **Heavy Naval and Air Use:** Specific examples such as heavy naval and air use in the region lead to significant disruptions in daily life, reinforcing the siege analogy.
- **Defense of Hamas's Actions:** The defense of Hamas's actions by leftist individuals and pro-Palestine supporters contributes to the perception of Gaza being under siege.
- **Comparisons to Historical Blockades:** Tweets draw parallels between Gaza's situation and historical blockades, such as Japan under US blockade during World War II and the Warsaw Ghetto, emphasizing the severity of the blockade.

These insights highlight the complex and emotionally charged nature of the discourse, showing how deeply ingrained sentiments and perceptions can drive extreme viewpoints. Understanding these causal links is crucial for modern use cases such as conflict monitoring, media analysis, and policy-making. By identifying the root causes of such descriptions, stakeholders can develop more targeted and effective strategies to address misinformation, reduce tensions, and promote a more balanced and informed public dialogue.

Chapter 8

Contributions

Our main contributions in this study are:

- A comprehensive KG framework has been developed to analyze causal relationships in datasets, enhancing our understanding of the underlying dynamics and sentiment propagation across multiple topics (**RQ1**) [52, 54].
- A novel approach to visualizing sentiment analysis through KGs has been introduced, enabling a detailed examination of entity relationships and their associated sentiments. This visualization aids in identifying key influencers and clusters within the dataset, particularly in the context of the Israel-Palestine conflict (**RQ1**) [8, 4].
- The methodology leveraged advanced natural language processing (NLP) techniques for entity extraction, relationship identification, and sentiment analysis, providing a robust framework for causal analysis in social media datasets. This includes the use of networkx and Node2Vec algorithms for building and visualizing the KG (**RQ1**) [9, 27, 19].
- A RAG model is proposed and validated, integrating KGs with LLMs to improve the accuracy and depth of causal reasoning. This hybrid approach bridges the gap between structured knowledge representation and advanced language understanding, enhancing the interpretability and actionable insights of the analysis (**RQ2**) [12, 13].
- The approach allowed for continuous updates and scalability, making it suitable for dynamically growing datasets. This ensures that the causal analysis remains

accurate as new data becomes available, as demonstrated in the COVID-19 tweets dataset (**RQ3**) [17].

- Qualitative and quantitative evaluations of the methods are conducted, demonstrating their effectiveness in providing deeper insights and improving decision-making processes in various contexts. This includes evaluations of centrality measures and connected components within the KGs (**RQ1**, **RQ2**, **RQ3**) [41, 64].

- A novel methodology to derive actionable insights from a time series analysis of a COVID-19 tweets dataset is proposed. This method facilitates the tracing of tweets that lead to specific causations, ensuring interoperability and transparency, which is currently lacking in LLM-only solutions (**RQ3**) [17, 4].

- The method can be applied to continuously growing datasets, not just static snapshots, providing real-time updates and insights (**RQ3**) [17].

- The ability to perform causal analysis is enhanced by integrating KG results and sentiment analysis into a broader framework that combines time series data with advanced ML models. This comprehensive approach enables a more holistic understanding of the data and the underlying causal relationships (**RQ3**) [52, 8].

- The contributions also include the development of a system that adapts to the dynamic nature of social media data, ensuring that insights remain current and relevant as new data is incorporated (**RQ3**) [17].

- Finally, a detailed examination of the effectiveness of our methodologies in identifying causal relationships is provided, demonstrating the potential of the approaches to enhance understanding and provide actionable insights from complex and evolving datasets (**RQ1**, **RQ2**, **RQ3**) [36, 38].

In the interest of fostering collaboration and advancing research, the code is made publicly available to the research community. This will allow other researchers to validate, replicate, and build upon my work, thereby contributing to the collective effort

to drive innovation and discovery in this field. The link to the GitHub repository is https://github.com/rahulst2010.

1

[1]https://github.com/rahulst2010

Chapter 9

Conclusion and Future Work

In this research, the effectiveness of integrating Knowledge Graphs (KGs) with Large Language Models (LLMs) using the Retrieval-Augmented Generation (RAG) method for causal analysis of social media data has been demonstrated. The proposed model, PRAGyan, showed promising results on the COVID-19 dataset, indicating its potential for various social network applications, such as identifying misinformation and key concerns during sensitive events. The framework significantly improves the quality of causal inferences, achieving an improvement in accuracy by 10% compared to the baseline model.

In the future, several enhancements are planned to further improve and apply the proposed model:

1. **Refinement of Knowledge Graph Construction**: Enhancing the KG by including both causation and correlation information to ensure a better retrieval of context for LLM queries [36]. This improvement will provide a more robust understanding of causal relationships in the data.

2. **Exploration of Advanced Embedding Techniques**: Investigating alternative embedding techniques such as GraphSAGE [41] or Graph Attention Networks (GAT) [42] to capture richer semantic information. These techniques can improve the model's ability to understand and represent complex relationships within the data.

3. **Deploying a Tool**: Developing and deploying a tool with a user-friendly interface that captures data in real-time, continuously updates a KG using Neo4j, and performs causal analysis based on user queries. It will feature advanced visualization features, such as interactive dashboards created with D3.js and Tableau, to make the insights generated by the model more accessible and actionable for decision-

makers [65]. WebSockets for real-time data integration [66] and RESTful APIs for seamless interaction between the front-end and back-end will to used to ensure the model's responsiveness [67]. This integration of LLMs and dynamic KG updates will provide a practical, efficient solution for various applications to make it easier for users to gain valuable insights quickly.

4. **Applicability to Other Problems**: Extending the application of the model to various domains through social media, such as healthcare, finance, and disaster management. For instance:

    - In healthcare, analyzing patient feedback to identify common concerns and misinformation about treatments.
    - In finance, monitoring market sentiment and identifying potential misinformation affecting stock prices.
    - In disaster management, real-time analysis of social media to identify urgent needs and misinformation during natural disaster events.

By implementing these enhancements and exploring new applications, the proposed model can become a versatile and powerful tool for causal analysis across different domains, providing valuable insights for informed decision-making.

REFERENCES

[1] A. Morrison. (Nov 11, 2021) Data intelligibility and the quest for mutually understood meaning. [Online]. Available: https://www.datasciencecentral.com/data-intelligibility-and-the-quest-for-mutually-understood/

[2] G. Deshpande *et al.*, "User feedback from tweets vs app store reviews: An exploratory study of frequency, timing and content," in *2018 5th international workshop on artificial intelligence for requirements engineering (AIRE)*. IEEE, 2018, pp. 15–21.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] F. Shamrat *et al.*, "Sentiment analysis on twitter tweets about covid-19 vaccines using nlp and supervised knn classification algorithm," *Indonesian Jour. of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 463–470, 2021.

[5] M. Kejriwal *et al.*, *Knowledge graphs: Fundamentals, techniques, and applications*. MIT Press, 2021.

[6] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic web*, vol. 8, no. 3, pp. 489–508, 2017.

[7] C. Peng *et al.*, "Knowledge graphs: Opportunities and challenges," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13 071–13 102, 2023.

[8] U. Khurshid and I. Ihsan, "Knowledge graph embedding based sentiment analysis of product reviews using lstm and fuzzy logic," *Journal of Computing & Biomedical Informatics*, vol. 5, no. 01, pp. 240–242, 2023.

[9] X. Liu *et al.*, "Recognizing named entities in tweets," in *Proc. of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 359–367.

[10] B. Haradizadeh *et al.*, "Tweeki: Linking named entities on twitter to a knowledge graph, 2020," in *Workshop on Noisy Usergenerated Text*.

[11] J. Vizcarra, K. Kozaki, M. Torres Ruiz, and R. Quintero, "Knowledge-based sentiment analysis and visualization on social networks," *New Generation Computing*, vol. 39, pp. 199–229, 2021.

[12] P. Lewis, E. Perez *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[13] Y. Ding, W. Fan *et al.*, "A survey on rag meets llms: Towards retrieval-augmented large language models," *arXiv preprint arXiv:2405.06211*, 2024.

[14] C. Chan, C. Jiayang *et al.*, "Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations," in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 684–721.

[15] S. Gopalakrishnan *et al.*, "Text to causal knowledge graph: A framework to synthesize knowledge from unstructured business texts into causal graphs," *Information*, vol. 14, no. 7, p. 367, 2023.

[16] D. Dessì *et al.*, "Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain," *Future Generation Computer Systems*, vol. 116, pp. 253–264, 2021.

[17] G. Preda, "Covid19 tweets (2020)," *Available in: https://www. kaggle. com/gpreda/covid19-tweets. Accesses on May*, vol. 21, 2021.

[18] H. Kayesh, M. S. Islam, and J. Wang, "On event causality detection in tweets," *arXiv preprint arXiv:1901.03526*, 2019.

[19] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[20] C. Tjortjis, *Graph Databases: Applications on Social Media Analytics and Smart Cities.* CRC Press, 2023.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, pp. 2825–2830, 2011.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[23] S. Storks, Q. Gao, and J. Y. Chai, "Recent advances in natural language inference: A survey of benchmarks, resources, and approaches," *arXiv preprint arXiv:1904.01172*, 2019.

[24] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier *et al.*, "Knowledge graphs," *ACM Computing Surveys (Csur)*, vol. 54, no. 4, pp. 1–37, 2021.

[25] Y. Wang, L. Dong *et al.*, "Kg2vec: A node2vec-based vectorization model for knowledge graph," *Plos one*, vol. 16, no. 3, p. e0248552, 2021.

[26] X. Liu *et al.*, "Named entity recognition for tweets," *ACM Tran. on Intelligent Systems and Technology (TIST)*, vol. 4, no. 1, pp. 1–15, 2013.

[27] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. of the 22nd ACM SIGKDD Inter. Conf. on Knowledge discovery and data mining*, 2016, pp. 855–864.

[28] S. Hong, M. Wu *et al.*, "Event2vec: Learning representations of events on temporal sequences," in *Web and Big Data: First Inter. Joint Conf., APWeb-WAIM 2017, Beijing, China, July 7–9, 2017, Proc., Part II 1*. Springer, 2017, pp. 33–47.

[29] S. O. T. D. Science. (Apr 26, 2024) Relation extraction with llama3 models. [Online]. Available: https://towardsdatascience.com/relation-extraction-with-llama3-models-f8bc41858b9e

[30] H. Yu, A. Gan *et al.*, "Evaluation of retrieval-augmented generation: A survey," *arXiv preprint arXiv:2405.07437*, 2024.

[31] Z. Jin, Y. Chen *et al.*, "Cladder: Assessing causal reasoning in language models," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[32] F. A. Tan and D. o. Paul, "Constructing and interpreting causal knowledge graphs from news," in *Proceedings of the AAAI Symposium Series*, vol. 1, no. 1, 2023, pp. 52–59.

[33] L. Ehrlinger and W. WökJ.kg^{-1}K^{-1}, "Towards a definition of knowledge graphs." *SEMANTiCS (Posters, Demos, SuCCESS)*, vol. 48, no. 1-4, p. 2, 2016.

[34] U. Jaimini and A. Sheth, "Causalkg: Causal knowledge graph explainability using interventional and counterfactual reasoning," *IEEE Internet Computing*, vol. 26, no. 1, pp. 43–50, 2022.

[35] E. Kıcıman, R. Ness *et al.*, "Causal reasoning and large language models: Opening a new frontier for causality," *arXiv preprint arXiv:2305.00050*, 2023.

[36] Z. Jin, J. Liu *et al.*, "Can large language models infer causation from correlation?" *arXiv preprint arXiv:2306.05836*, 2023.

[37] D. Loureiro *et al.*, "Timelms: Diachronic language models from twitter," *arXiv preprint arXiv:2202.03829*, 2022.

[38] L. Dieudonat, K. Han *et al.*, "Exploring the combination of contextual word embeddings and knowledge graph embeddings," *arXiv preprint arXiv:2004.08371*, 2020.

[39] S. Pan, L. Luo *et al.*, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[40] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[41] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[42] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.

[43] T. Wu, A. Khan *et al.*, "Efficiently embedding dynamic knowledge graphs," *Knowledge-Based Systems*, vol. 250, p. 109124, 2022.

[44] Z. Lu *et al.*, "Vgcn-bert: augmenting bert with graph embedding for text classification," in *Advances in Information Retrieval: 42nd European Conf. on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proc., Part I 42*. Springer, 2020, pp. 369–382.

[45] J. M. Mooij, J. Peters, D. Janzing *et al.*, "Distinguishing cause from effect using observational data: methods and benchmarks," *Journal of Machine Learning Research*, vol. 17, no. 32, pp. 1–102, 2016.

[46] Y. Gao, Y. Xiong *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[47] W. X. Zhao, K. Zhou, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[48] S. Minaee, T. Mikolov *et al.*, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

[49] T. Mikolov, K. Chen *et al.*, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[50] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[51] R. Angles and C. Gutierrez, "Survey of graph database models," *ACM Computing Surveys (CSUR)*, vol. 40, no. 1, pp. 1–39, 2008.

[52] Z. Xu and Y. Dang, "Data-driven causal knowledge graph construction for root cause analysis in quality problem solving," *International Journal of Production Research*, vol. 61, no. 10, pp. 3227–3245, 2023.

[53] S. Wadhwa *et al.*, "Revisiting relation extraction in the era of large language models," in *Proc. of the Conf.. Association for Computational Linguistics. Meeting*, vol. 2023. NIH Public Access, pp. 1:55–66.

[54] L. Du, X. Ding *et al.*, "e-care: a new dataset for exploring explainable causal reasoning," *arXiv preprint arXiv:2205.05849*, 2022.

[55] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," *arXiv preprint arXiv:2309.15217*, 2023.

[56] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[57] N. Alqaydeh, "Named entity recognition (ner) corpus," 2024, accessed: 2024-07-07. [Online]. Available: https://www.kaggle.com/datasets/naseralqaydeh/named-entity-recognition-ner-corpus

[58] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, L. Yu, Y. Liu, J. Li, B. Xiong, D. Xiong *et al.*, "Evaluating large language models: A comprehensive survey," *arXiv preprint arXiv:2310.19736*, 2023.

[59] Asaniczka, "Daily public opinion on israel-palestine war," 2024. [Online]. Available: https://www.kaggle.com/dsv/9026277

[60] A. Kokeyo, "Exploring the dynamics of social media in shaping narratives and perceptions in the israeli-palestinian conflict: preliminary reflections," *African Journal of Emerging Issues*, vol. 5, no. 17, pp. 181–194, 2023.

[61] L. Lehrs, "Interlocking peace processes: Between competing and complementing peacemaking efforts in interlocking conflicts," *Cooperation and Conflict*, vol. 58, no. 4, pp. 485–501, 2023.

[62] P. N. Howard and M. M. Hussain, *Democracy's fourth wave?: digital media and the Arab Spring*. Oxford University Press, 2013.

[63] Y. David and N. Shalhoub-Kevorkian, "Racializing human rights: political orientation, racial beliefs, and media use as predictors of support for human rights violations–a case study of the israeli-palestinian conflict," *Ethnic and Racial Studies*, vol. 46, no. 10, pp. 1947–1971, 2023.

[64] K. Papineni *et al.*, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[65] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[66] I. Fette and A. Melnikov, "The websocket protocol," in *IETF RFC 6455*, 2011, pp. 1–50.

[67] L. Richardson, M. Amundsen, and S. Ruby, *RESTful Web APIs*. O'Reilly Media, Inc., 2013.

APPENDICES

A Algorithm for pRAGyan

Algorithm 1: Pseudo Code for Causal Inference with Knowledge Graph Enhanced LLM

Input: Text data T, Relation extraction model R, Node2Vec model N, BERT Uncased Encoder B, LLM M, Graph database DB
Output: Causal inferences
Input: Text data T, Relation extraction model R, Node2Vec model N, LLM M, Graph database DB
Output: Causal inferences

$T_{preprocessed} \leftarrow$ Preprocess(T)
$E, Rel \leftarrow$ ExtractEntitiesAndRelations($T_{preprocessed}, R$)
$G \leftarrow$ ConstructKnowledgeGraph(E, Rel)
if GraphDBExists(DB) **then**
  for each node, edge in G **do**
    AddNodeToGraphDB(node, DB)
    AddEdgeToGraphDB(edge, DB)
  end for
else
  InitializeGraphDB(G, DB)
end if
$E_{embeddings} \leftarrow$ GenerateEmbeddings(E, N)
for each *query* in *queries* **do**
  $query\_embedding \leftarrow$ GenerateEncoding($query, B$)
  $similarities \leftarrow$ CalculateSimilarities($query\_encoding, E_{embeddings}$)
  $TopK \leftarrow$ SelectTopKSubgraph($G, similarities, 5$)
  $context \leftarrow$ RetrieveContext($TopK$)
  Using Retrieved Context (RAG)
  $Answer\_RAG \leftarrow$ GenerateAnswer($query, context, M$)
  StoreAnswer(Answer_RAG)
  Using Entire Corpus
  $Answer\_Corpus \leftarrow$ GenerateAnswer($query, T_{preprocessed}, M$)
  StoreAnswer(Answer_Corpus)
end for
$F1_{RAG} \leftarrow$ EvaluateModel($Answers\_RAG$, TestingDataset)
$F1_{Corpus} \leftarrow$ EvaluateModel($Answers\_Corpus$, TestingDataset)
$\Delta F1 \leftarrow F1_{RAG} - F1_{Corpus}$
for each *query* in *queries* **do**
  $Best\_Answer \leftarrow$ SelectBestAnswer($Answer\_RAG, Answer\_Corpus$)
  StoreBestAnswer(Best_Answer)
end for
$CausalInferences \leftarrow$ InferCauses($Best\_Answers$)
return $CausalInferences$