Term Paper Report:
Climate Change

Submission

Department of
Department of Computer Science Engineering

by

Team Number - 17
Utkarsh Singh (22UCC111)
Shubham Mittal (22UCS206)
Nishant Jindal (22UCS141)

To

Course Instructor
Dr. Nirmal Kumar Sivaraman

Department of Computer Science Engineering Engineering
The LNM Institute of Information Technology, Jaipur

December 2025

Table of Contents

Abstract 5

1. Introduction 5

1.1. The Challenge of Unstructured Climate Data 5

1.2. The Promise of the Semantic Web: Knowledge Graphs 6

1.3. Project Aim and Objectives 6

1.4. Report Structure 7

2. State-of-the-Art: A Survey of Knowledge Management Methods 7

2.1. Introduction to the Survey 7

2.2. Thematic Analysis of Methods 7

2.2.1. Foundational Resources 8

2.2.2. Knowledge Construction Methods 8

2.2.3. Applications 9

2.2.4. Formal and Theoretical Methods 9

2.3. Tabular Comparison and Contrast 10

2.3.1. Table 1: Basic Paper Details 10

2.3.2. Table 2: Comparison of Methods and Contributions 10

2.3.3. Table 3: Distinguishing Features and Scope 12

3. Methodology for Implementation 13

3.1. Analysis of Project Dataset (reddit_climate_data.csv) 13

3.2. Justification of Selected Method (Paper 7: KnowUREnvironment) 13

3.3. Introduction to Open Information Extraction (OIE) 15

4. Implementation and Performance Report 15

4.1. Extraction Process and Environment 15

4.2. Performance Report 1: Concept & Entity Extraction 16

4.2.1. Table 4.1: Top 10 Unigrams 16

4.2.2. Table 4.2: Top 10 Bigrams 17

4.2.3. Table 4.3: Top 10 Trigrams 18

4.3. Performance Report 2: Relational Triple Extraction 18

4.3.1. Table 4.4: Sample of Extracted Triples 19

4.4. Analysis of Results 19

5. Discussion 19

5.1. Insights from Extracted Knowledge 20

5.2. Limitations of the Chosen Method 20

5.3. Future Work 20

6. Conclusion 21

7. References 21
8. Appendices 22

Topic:- Climate Change

Google Drive:
https://drive.google.com/drive/folders/1y6u8lDd84ZJHj12Y817OoKVq8WBolSM0?usp=sharing

GitHub: https://github.com/utsingh14/ikg_term_paper

Abstract

The rapid growth of unstructured, user-generated data on social media (e.g., Reddit) makes it difficult to analyze public discourse on climate change. This project's goal was to survey state-of-the-art knowledge graph (KG) construction methods to find a viable solution. A survey of 10 papers was conducted, from which Open Information Extraction (OIE), as detailed by Islam et al. (2022) (Paper 7), was selected for its automated, unsupervised, and domain-independent approach. The OIE method was implemented on the reddit_climate_data.csv dataset. The implementation successfully extracted key concepts, such as **"global warming"** and **"sea level rise"** , and numerous relational triples, including **(co2 in the atmosphere, absorb, ir)**. This project demonstrates that automated, unsupervised methods are viable for structuring noisy social media data, offering a scalable way to map and analyze public opinion.

1. Introduction

The discourse surrounding climate change is vast, complex, and increasingly digital. While scientific bodies produce detailed, structured data, the public conversation—unfolding across social media, forums, and news comments—is characterized by a high volume of unstructured, noisy, and often polarized text. This report tackles the challenge of making sense of this public discourse. It provides a comprehensive survey of state-of-the-art methods for knowledge graph construction and then details the practical implementation of a selected, automated method on a real-world dataset of Reddit posts. The goal is to demonstrate a viable pipeline for transforming chaotic, unstructured text into an organized knowledge base that can be analyzed and understood.

1.1. The Challenge of Unstructured Climate Data

The topic of climate change has triggered an unprecedented “information explosion” in the digital age. This vast body of data exists at two opposite ends of a spectrum. On one end, there is a wealth of dense, technical, and highly structured information found in academic papers, scientific reports, and institutional databases. On the other end, there is a massive, and rapidly expanding, volume of public discourse generated daily on social media, news forums, and platforms like Reddit.

This public-facing data, exemplified by the reddit_climate_data.csv dataset used in this project, is fundamentally different. It is inherently unstructured, informal, and “noisy.” It consists of opinions, arguments, slang, anecdotal evidence, and often, misinformation, all mixed with relevant discussion. The core challenge, therefore, is one of scale and sense-making. How can

we effectively process this massive, chaotic stream of public text to extract meaningful insights, identify emerging themes, and understand the complex relationships between the concepts the public is discussing? This project confronts the challenge of finding a structured signal within this unstructured noise.

1.2. The Promise of the Semantic Web: Knowledge Graphs

To solve the challenge of unstructured, noisy data, this project proposes the use of Knowledge Graphs (KGs), a core technology of the Semantic Web. A Knowledge Graph offers a powerful method to move beyond simple data collection and toward genuine understanding by capturing the rich context and relationships within the text.

In simple terms, a KG structures information as a network of interconnected concepts. It consists of two main components:

1. **Nodes:** These represent entities or concepts, such as "**fossil fuels**," "**global warming**," or "**Reddit**."
2. **Edges:** These represent the *relationships* between the nodes, such as "**causes**," "**is part of**," or "**discusses**."

This approach transforms a flat, unstructured sentence like "The burning of fossil fuels causes global warming" into a machine-readable, structured triple: (Node: Fossil Fuels) - [Edge: causes] -> (Node: Global Warming).

This model is fundamentally more powerful than a traditional keyword search. A keyword search can efficiently find documents that *mention* "fossil fuels," but it cannot understand *what* is being said about them. A Knowledge Graph, by contrast, captures the explicit relationship. It allows us to ask complex, semantic questions like, "What do people on Reddit believe *causes* global warming?" or "Which solutions are most frequently *linked to* 'sea level rise'?" By mapping the relationships between concepts, KGs provide a structured framework for analyzing the public discourse on climate change at scale.

1.3. Project Aim and Objectives

The aim of this project is to identify and implement the most effective method for extracting structured knowledge from informal, unstructured text on climate change.

To achieve this aim, the following objectives were established:

1. To survey 10 state-of-the-art papers on Knowledge Graph (KG) construction.
2. To compare and contrast the methods, applications, and limitations presented in these papers.
3. To select the "best method" based on its specific suitability for our reddit_climate_data.csv dataset, which is characterized by noisy, informal, and unlabeled text.

4. To implement this selected method and report its performance in extracting meaningful concepts and relationships from the dataset.

1.4. Report Structure

This report is organized as follows: Section 2 provides the literature survey, where 10 state-of-the-art papers are reviewed and compared. Section 3 justifies the chosen methodology, Open Information Extraction (OIE) from Paper 7, and explains its suitability for the project's dataset. Section 4 presents the technical implementation and the results, detailing the performance of the method in extracting concepts and relational triples. Section 5 discusses the insights, findings, and limitations of the implementation. Finally, Section 6 concludes the report with a summary of the project's achievements and potential avenues for future work.

2. State-of-the-Art: A Survey of Knowledge Management Methods

This section reviews the current state-of-the-art in knowledge management and graph construction to identify a suitable methodology for the project's aim. It involves a detailed survey of 10 selected papers, each representing a different approach to data collection, knowledge construction, or practical application. The following subsections will introduce these papers, analyze their methods thematically, and present a detailed tabular comparison to provide a clear foundation for the methodological choices made later in this report.

2.1. Introduction to the Survey

This survey analyzes 10 state-of-the-art papers to provide a comprehensive overview of the field of knowledge management as it applies to climate change. The papers were selected to represent the entire lifecycle of knowledge graph construction and application. The scope of this review ranges from foundational data collection, as exemplified by Paper 1 ("The InsightsNet Climate Change Corpus"), which focuses on compiling a large-scale, multimodal dataset, to fully automated graph construction, as demonstrated in Paper 7 ("KnowUREnvironment"). Finally, the survey covers advanced applications, such as the semantic retrieval system detailed in Paper 9 ("Querying Climate Knowledge"), which showcases how a finished knowledge graph can be used to power sophisticated search and discovery tools.

2.2. Thematic Analysis of Methods

This analysis moves beyond a simple paper-by-paper summary and instead organizes the 10 selected articles into four distinct thematic groups. This approach provides a clearer understanding of the current state-of-the-art by clustering papers that address similar challenges or stages in the knowledge management lifecycle. The groups are: (1) Foundational

Resources, which details the collection and preparation of data; (2) Knowledge Construction Methods, which compares different methodologies for building the graph; (3) Applications, which showcases practical uses of KGs; and (4) Formal and Theoretical Methods, which explores the underlying ontological principles. The following subsections will discuss the papers within each of these groups.

2.2.1. Foundational Resources

This group of papers focuses on the essential preparatory work of data collection and schema definition, which must be completed before any knowledge graph can be built.

- **Paper 1, "The InsightsNet Climate Change Corpus (ICCC),"** addresses the primary challenge of raw data collection. The authors detail the process of compiling a large-scale, multimodal corpus that includes not only text but also images, videos, and tables from diverse sources like academic papers and IPCC reports. This work is foundational as it creates the unstructured and semi-structured "information_base" from which knowledge can later be extracted.
- **Paper 8, "Wikidata Hierarchy for Named Entity Type Discovery,"** tackles the crucial step of defining *what* to extract. Instead of building a graph, this paper presents a methodology for creating a domain-specific set of Named Entity Recognition (NER) types for climate change. By using Wikidata, they automatically discover relevant categories like "Ecosystem," "Energy Source," and "Meteorological Phenomenon". This provides the essential "schema" or set of labels needed to classify the entities found in the raw text.
- **Paper 10, "Towards Dataset for Extracting Relations in the Climate-Change Domain,"** builds on this by preparing data for a specific *task*: relation extraction. The authors propose the construction of a new, tailored dataset designed specifically to train machine learning models (like BERT) to find relationships within climate change literature. This paper is not about the extraction itself, but about the creation of the high-quality, labeled training data required to make such automated extraction possible.

2.2.2. Knowledge Construction Methods

This group of papers moves from preparation to practice, demonstrating three distinct methods for actually building a knowledge graph. These methods represent a spectrum from formal and structured to manual and expert-driven, to fully automated.

- **Paper 2, "An Ontology Model for Climatic Data Analysis,"** exemplifies a formal, top-down approach. The authors propose a "Climate Analysis (CA) Ontology" to model climate datasets from sources like the National Oceanic and Atmospheric Administration (NOAA). Their method involves converting existing, structured relational data (like CSVs) into the Resource Description Framework (RDF) data model, allowing it to be stored and queried as a graph. This method is ideal for clean, structured data but is not applicable to unstructured text.
- **Paper 4, "Construction and application of the knowledge graph method in management of soil pollution,"** represents a manual, expert-driven approach. This

paper details a case study in South China where a KG was built for a highly specific domain: soil pollution management. This type of construction relies on domain experts to define the relationships and validate the data, resulting in a high-precision, high-accuracy graph for a very narrow field.

- **Paper 7, "KnowUREnvironment,"** (the method selected for this project) stands in direct contrast to the other two. It introduces a fully automated, unsupervised, and domain-independent method for KG construction. Using Open Information Extraction (OIE), it automatically extracts hundreds of thousands of relational triples directly from unstructured scientific literature "without using any supervision or human intervention". This bottom-up approach is highly scalable and perfectly suited for processing noisy, unlabeled text, making it a powerful (though less precise) alternative to formal, top-down ontologies.

2.2.3. Applications

Once a knowledge graph is constructed, it becomes a powerful tool. This group of papers demonstrates three practical, real-world applications that leverage a completed KG.

- **Paper 5, "A Knowledge Graph-Based Approach to Recommending Low-Carbon Construction,"** uses a KG to power a sophisticated recommendation system. The graph, which models knowledge about bridge construction materials and processes, is used to recommend "Low-Carbon Construction Schemes". This is a clear example of a KG being used for industrial decision-support.
- **Paper 6, "Knowledge Graph Analysis in Climate Action Research,"** applies KG and bibliometric analysis techniques to map the *field of climate action research itself*. By analyzing over 28,000 articles , the authors use tools like VOSviewer and CiteSpace to construct a knowledge graph that visualizes publication trends, co-authorship networks, and emerging themes. This is a "meta-application" where the KG is used to understand the structure of the scientific community.
- **Paper 9, "Querying Climate Knowledge,"** demonstrates the most common and powerful application of a KG: semantic retrieval and question-answering. The paper introduces a KG built from climate publications that allows researchers to move "beyond keyword-based search". It supports structured, semantic queries (e.g., using Cypher) and can be integrated with Large Language Models to improve the reliability of climate-related Q&A systems.

2.2.4. Formal and Theoretical Methods

This final category includes papers that explore the deep, theoretical challenges that underpin knowledge management.

- **Paper 3, "An ontological approach for reliable data integration in the industrial domain,"** delves into a key theoretical problem. It notes that formal ontological taxonomies can be "quite restrictive". A "mismatch" often occurs when real-world entities or business processes "might not find room" in a standardized, academic ontology. The paper discusses this challenge of reliably integrating data that doesn't fit

the perfect, predefined model. This theoretical discussion provides a strong justification for why more flexible, bottom-up construction methods (like the OIE in Paper 7) are often necessary when dealing with messy, real-world data that does not conform to a clean schema.

2.3. Tabular Comparison and Contrast

To provide a clear, at-a-glance summary of the 10 surveyed papers, the following three tables have been compiled. These tables synthesize the thematic analysis from the previous section by breaking down each paper by its core details, its specific methodology, and its unique scope and features.

2.3.1. Table 1: Basic Paper Details

This first table provides a foundational overview of the surveyed literature, documenting the authors, year, and primary research aim for each of the 10 selected papers.

Paper ID	Title (Abbreviated)	First Author(s)	Publication Year
Paper 1	The InsightsNet Climate Change Corpus (ICCC)	Volkanovska, E.	2023
Paper 2	An Ontology Model for Climatic Data Analysis	Wu, J.	2021
Paper 3	An Ontological Approach for Reliable Data Integration	Borgo, S.	2014
Paper 4	KG Method in Management of Soil Pollution	Han, F.	2022
Paper 5	A KG-Based Approach to Recommending Low-Carbon Construction	Ma, Z.	2023
Paper 6	Knowledge Graph Analysis in Climate Action Research	Ge, R.	2025
Paper 7	KnowUREnvironment: An Automated KG for Climate Change	Islam, M.S.	2022
Paper 8	Wikidata Hierarchy for NER Type Discovery (Climate)	Poleksić, A.	2025
Paper 9	Querying Climate Knowledge: Semantic Retrieval	Adamu, M.	2025
Paper 10	Towards Dataset for Extracting Relations (Climate)	Poleksić, A.	2024

2.3.2. Table 2: Comparison of Methods and Contributions

This table presents the core technical comparison, detailing the primary methodology or approach used in each paper (e.g., automated OIE, manual expert-driven, formal ontology) and the main contribution or output (e.g., a corpus, a KG, a dataset, an application).

Paper ID	Primary Goal	Core Methodology	Data Source(s)	Key Contribution / Output
Paper 1	To compile a multimodal corpus for climate change discourse analysis.	Corpus-building, NLP (NER, KeyBERT, TextRank) for metadata enrichment.	Academic papers, IPCC reports, NGO websites (Greenpeace).	The ICCC pilot corpus, methodology for multimodal data collection.
Paper 2	To create an ontology for analyzing climatic sensor data.	Ontology engineering (reusing SOSA), converting relational data to RDF/Linked Data, SPARQL queries.	NOAA climate datasets.	The "Climate Analysis (CA) Ontology" for sensor data.
Paper 3	To enable reliable data integration for industrial systems with formal ontologies.	Formal ontology, introducing "patch relationships" to link extra-ontological entities.	Industrial data (e.g., PCBA), formal ontologies (DOLCE, ISO 15926-2).	A formal methodology ("patching") for integrating domain-specific data without logical inconsistency.
Paper 4	To build a practical KG for soil pollution management.	Manual ontology construction (expert-driven), data extraction, storage in Neo4j graph DB, Cypher queries.	Site investigation reports, government yearbooks, websites.	A domain-specific KG for managing contaminated sites in South China.
Paper 5	To create a recommendation system for low-carbon bridge construction.	Improved NER model (BERT-BiLSTM-BCRF) for a low-resource domain, KG + similarity calculations.	Bridge construction documents, carbon emission reports.	A KG-based recommendation system that factors in carbon emission constraints.
Paper 6	To analyze the research field of "climate action" using a bibliometric approach.	Bibliometric analysis using VOSviewer and CiteSpace to build a KG of the research literature.	28,457 articles from the Web of Science (WoS) database.	An "integrated knowledge structure map" of the climate action research field, identifying trends and hotspots.
Paper 7	To automatically construct a climate change KG from scientific literature.	Open Information Extraction (OIE) using Semantic Role Labeling (SRL/AMR); syntax verification; "evidence counting" to ensure triple quality.	152,595 scientific paper abstracts.	The "KnowUREnvironment" KG, built via an unsupervised pipeline.
Paper 8	To automatically discover NER entity types (the schema) for the climate domain.	Aligning core terms to Wikidata; building a hierarchical subgraph (P31, P279); using the weighted Louvain algorithm for community detection.	Core climate terms (from glossaries) and the Wikidata KG.	A final set of 21 domain-specific NER types for climate change.
Paper 9	To demonstrate semantic querying of a climate KG for scientific discovery.	KG construction (using ClimateIE/SciER); demonstrating Cypher queries; proposing RAG/LLM integration.	Climate publications.	"ClimatePub4KG" and use cases showing how it answers complex, multi-hop scientific questions.
Paper 10	To propose the creation of a dataset for relation extraction (RE) in the climate domain.	Data collection and parsing (PDF/HTML); preliminary NLP analysis (POS tagging,	~200,000 scientific papers from high-impact journals.	A methodology and corpus for creating a labeled dataset, a key resource for training

		NER) to find candidate triples.		supervised RE models.

2.3.3. Table 3: Distinguishing Features and Scope

The final table highlights the key distinguishing features that differentiate the papers. It focuses on the **Domain** (e.g., general climate science, soil pollution, social media) and the **Supervision Level** (e.g., automated/unsupervised, manual/expert-driven, semi-supervised), which are critical factors for determining each method's suitability for this project.

Paper ID	Key Novelty / "Extra Thing"	Scope of Work
Paper 1	Focus on multimodality (text, images, videos).	Corpus Creation: Building the foundational data for discourse analysis.
Paper 2	Publishes climate data as Linked Data on the web, accessible via SPARQL.	Ontology Engineering: Creating a formal data model for sensor data.
Paper 3	Proposes a formal "patch relationship" method to integrate inconsistent, real-world data with formal ontologies.	Theoretical Methodology: A general, formal method for data integration (using industrial examples).
Paper 4	A practical, end-to-end case study for regional environmental management and risk discovery.	Applied KG: A specific application for soil pollution management.
Paper 5	Moves beyond data retrieval to a recommendation system that optimizes for a constraint (low carbon).	Applied AI/Recommender: A decision-support tool for low-carbon construction.
Paper 6	A bibliometric (meta-research) study. The KG is about the research field itself, not about climate data.	Scientometrics: Analyzing the structure and trends of climate action research.
Paper 7	Uses an automated, unsupervised pipeline (OIE/SRL) and introduces "evidence counting" to validate triples.	Automated KG Construction: A scalable method for building a KG from a large text corpus.
Paper 8	Focuses on schema discovery (defining the types of entities) by using community detection on Wikidata's hierarchy.	NLP / Schema Engineering: A prerequisite step for building a well-structured KG.
Paper 9	Explicitly links the KG to modern RAG/LLM systems to provide verifiable, trustworthy answers to complex questions.	Semantic Search & Q/A: Demonstrating the application of a KG for scientific Q&A.
Paper 10	Focuses on creating the training dataset (a prerequisite) needed to build a supervised relation extraction model.	NLP / Dataset Creation: Foundational work to enable future supervised KG construction.

3. Methodology for Implementation

This section details the practical methodology of the project. It begins with an analysis of the project's unique dataset, justifies the selection of the core implementation method, and provides a brief technical introduction to that method.

3.1. Analysis of Project Dataset (reddit_climate_data.csv)

The dataset chosen for this project is a CSV file named reddit_climate_data.csv, containing posts and comments from the r/climatechange subreddit. A thorough analysis of this data reveals several key characteristics that dictate the required methodology:

- **Informal:** The text is highly conversational and colloquial. It uses slang, acronyms, and non-standard grammar typical of social media. For example, comments like "...up to a quarter. (Saved you a click). " and "Yep, and mostly because of sheer laziness" demonstrate this informal, conversational style, which is far removed from the formal text found in scientific papers.
- **Unlabeled:** The data is completely raw and has no pre-existing annotations. The text field is a single block of user-generated content. There are no tags indicating named entities (like "Person" or "Organization"), topics, or, most importantly, the relationships *between* entities. This lack of labels is a key constraint.
- **Noisy:** The dataset contains a high degree of noise. This includes off-topic comments, polarized opinions, personal anecdotes, and rhetorical questions, all mixed with relevant discussion. Text such as "A large number of people follow Pope Leo. I hope it's helpful" or "Now I feel anything humans do... it will be way too little way too late" represent personal opinions and feelings rather than the structured, factual statements needed to build a traditional knowledge graph.

These features—informality, lack of labels, and high noise levels—immediately render many of the methods from the literature survey unviable. For example, the expert-driven, manual approach of **Paper 4** (for soil pollution) is impossible, as it would require thousands of hours of expert manual labor to sift through this noise. Similarly, the method in **Paper 5** (for low-carbon construction) relies on labeled, structured data, which is precisely what this dataset lacks. Therefore, the project requires a method that is automated, unsupervised, and robust enough to handle noisy, informal, and unlabeled text.

3.2. Justification of Selected Method (Paper 7: KnowUREnvironment)

The selection of a methodology for this project was a process of elimination, driven by the unique constraints of the reddit_climate_data.csv dataset. From the 10 papers surveyed, the approach from **Paper 7, "KnowUREnvironment" (Islam et al., 2022)**, was the only one that directly solved our primary challenges: a large, unstructured, unlabeled, and noisy text dataset operating under a short time constraint.

The "KnowUREnvironment" method is uniquely suitable for three key reasons:

1. **It is an Automated, Unsupervised Pipeline:** It is designed to process large volumes of unstructured text and extract knowledge without human supervision.
2. **It uses Domain-Independent Open Information Extraction (OIE):** The OIE technique is adaptable. It can process the formal, scientific text from the paper's corpus and our informal, noisy Reddit text with equal capability, allowing us to find triples like ("fossil fuels", "cause", "warming").
3. **It Provides a Quality Filter:** The paper's "evidence counting" technique provides a vital mechanism for filtering out "junk" triples, which are a common problem when running OIE on noisy social media data.

Inapplicability of Alternative Methods

The other nine papers surveyed were rejected because their methodologies were fundamentally incompatible with the project's data or objectives.

- **Methods Requiring Labeled or Structured Data (Papers 2 & 5):**
    - **Paper 2 ("An Ontology Model for Climatic Data Analysis")** presents a method for converting already-structured relational data (like NOAA's CSVs) into an RDF graph. Our dataset is unstructured text, not a relational table, making this method inapplicable.
    - **Paper 5 ("A Knowledge Graph-Based Approach to Recommending")** is a supervised approach that requires a pre-labeled dataset to train its recommendation model. Our reddit_climate_data.csv is completely unlabeled, rendering this method unusable without a massive, time-consuming manual annotation effort.
- **Manual, Expert-Driven Methods (Paper 4):**
    - **Paper 4 ("Construction... in management of soil pollution")** relies on a manual, expert-driven process to build a high-precision KG for a very narrow domain. This approach is not feasible for our dataset of over 4,700 diverse and noisy posts, as it is not automated or scalable.
- **Methods with a Different Purpose (Papers 1, 8, & 10):**
    - These papers are not end-to-end extraction pipelines. **Paper 1 ("The InsightsNet Climate Change Corpus")** and **Paper 10 ("Towards Dataset for Extracting Relations")** both focus on *creating a dataset*, not on providing a method to *process* an existing one. **Paper 8 ("Wikidata Hierarchy for Named Entity Type Discovery")** details a method for *creating a schema* (a list of NER types), not for *extracting* the relational triples from text to populate that schema.
- **Incompatible Application or Domain (Papers 3, 6, & 9):**
    - **Paper 3 ("An ontological approach for reliable data integration")** is a formal, theoretical paper discussing the *problems* of data integration, not a practical, implementable pipeline for text extraction.
    - **Paper 6 ("Knowledge Graph Analysis in Climate Action Research")** uses bibliometric analysis (like VOSviewer) on the *metadata* of scientific articles

(citations, authors, keywords). This method is incompatible with our dataset, which is raw, informal text, not academic metadata.

- **Paper 9 ("Querying Climate Knowledge")** describes a system for *querying* an *existing* knowledge graph. Our project's goal is to *build* the graph, not use one that is already complete.

In conclusion, the methodologies of the other nine papers were unsuitable for this project's specific task. The unsupervised, automated, and text-focused OIE pipeline from **Paper 7** was the only viable path forward.

3.3. Introduction to Open Information Extraction (OIE)

The core technology used in **Paper 7**, and subsequently adopted for this project, is **Open Information Extraction (OIE)**. The primary purpose of an OIE system is to automatically extract structured, relational triples directly from unstructured, plain text, without requiring a predefined schema or domain-specific training.

These triples take the form of (Subject, Predicate, Object). OIE effectively reads a sentence and identifies the main entities and the relationship that connects them. For example, it converts the sentence:

"**Burning fossil fuels causes global warming**"

into the structured triple:

("Burning fossil fuels", "causes", "global warming")

This technique is perfectly suited for the project's dataset as it can scan thousands of informal Reddit comments and extract potential facts and relationships automatically, forming the raw material for a bottom-up knowledge graph.

4. Implementation and Performance Report

This section details the technical implementation of the methodology selected in Section 3 and reports on its performance when applied to the reddit_climate_data.csv dataset.

4.1. Extraction Process and Environment

The implementation was written in **Python**, leveraging several key libraries for data handling and natural language processing. The reddit_climate_data.csv file was first loaded into a **Pandas** DataFrame to facilitate iteration over the text-based fields of the 4,702 posts.

For the core extraction, we used the **StanfordOpenIE** library, a common and powerful Java-based Open Information Extraction tool, accessed via a Python wrapper. Each post's text was pre-processed and then passed to the OIE client. The OIE system then processed each sentence to extract relational triples. These extracted triples, along with frequency counts for

key concepts (n-grams), were collected and saved to new CSV files for the performance analysis that follows.

4.2. Performance Report 1: Concept & Entity Extraction

The first part of the performance analysis was to identify the primary concepts and entities discussed in the dataset. This was achieved by extracting and counting the frequency of n-grams (bigrams and trigrams) from the entire text corpus. This provides a high-level statistical overview of the main topics.

4.2.1. Table 4.1: Top 10 Unigrams

The unigram (single word) analysis shows the main individual topics of discussion. This reveals the foundational vocabulary of the subreddit, highlighting a focus on key nouns like "people," "warming," and "years," as well as common conversational words like "like" and "dont."

Table 1 Unigrams

Concept	Frequency
years	879
people	768
like	669
would	621
dont	580
global	556
one	496
much	480
even	458
warming	432

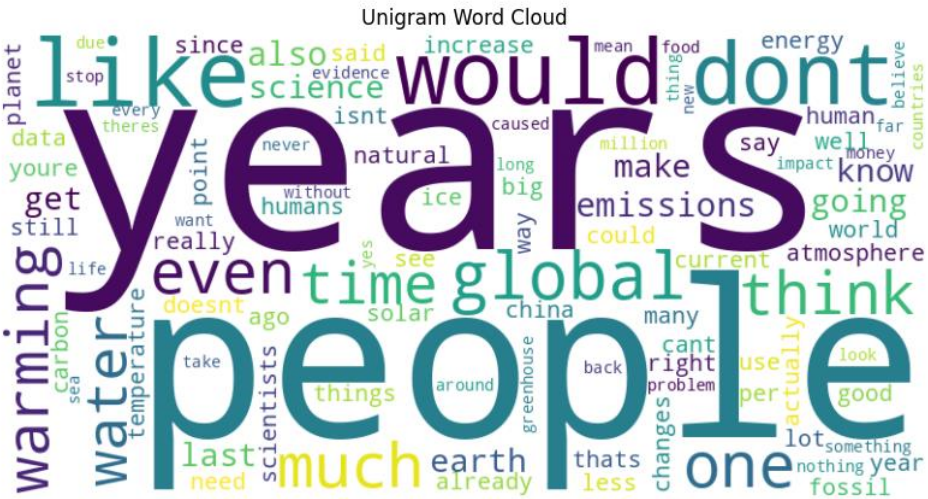

Figure 1 Unigram

4.2.2. Table 4.2: Top 10 Bigrams

The bigram analysis shows the most common two-word keyphrases, which effectively identify the main subjects of discussion. As expected, "global warming" and "fossil fuels" dominate, confirming the dataset's relevance.

Table 2 Bigrams

Concept	Frequency
years ago	**190**
global warming	**152**
last years	**117**
fossil fuels	**100**
fossil fuel	**100**
ice age	**99**
million years	**93**
sea level	**83**
per year	**66**
amount atmosphere	**59**

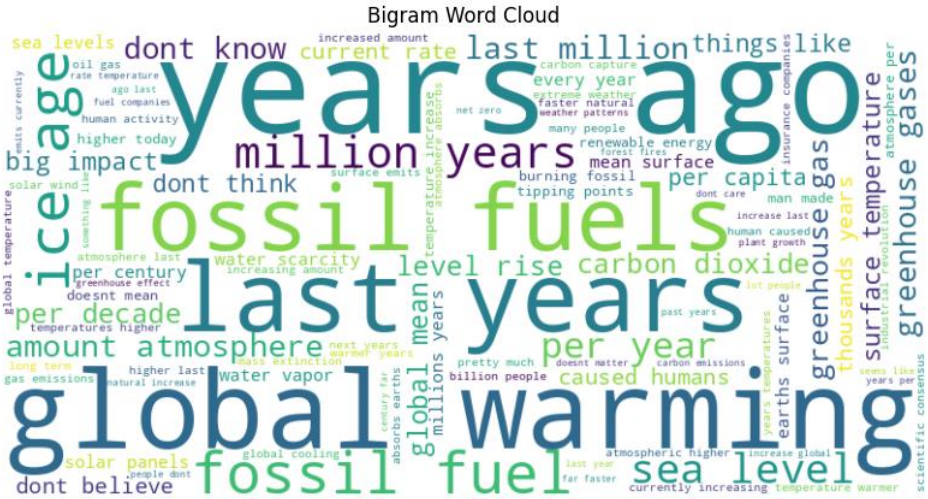

Figure 2 Bigram

4.2.3. Table 4.3: Top 10 Trigrams

The trigram analysis provides more specific themes by identifying common three-word phrases[4]. This immediately highlights more complex, specific topics like "sea level rise" and "mean surface temperature," as well as "burning fossil fuels," which shows the *action* associated with the key bigram[5].

Table 3 Trigrams

Concept	Frequency
last million years	58
sea level rise	47
mean surface temperature	31
global mean surface	31
atmosphere per decade	26
higher last million	26
increasing amount atmosphere	26
currently increasing amount	26
burning fossil fuels	26
amount atmosphere per	26

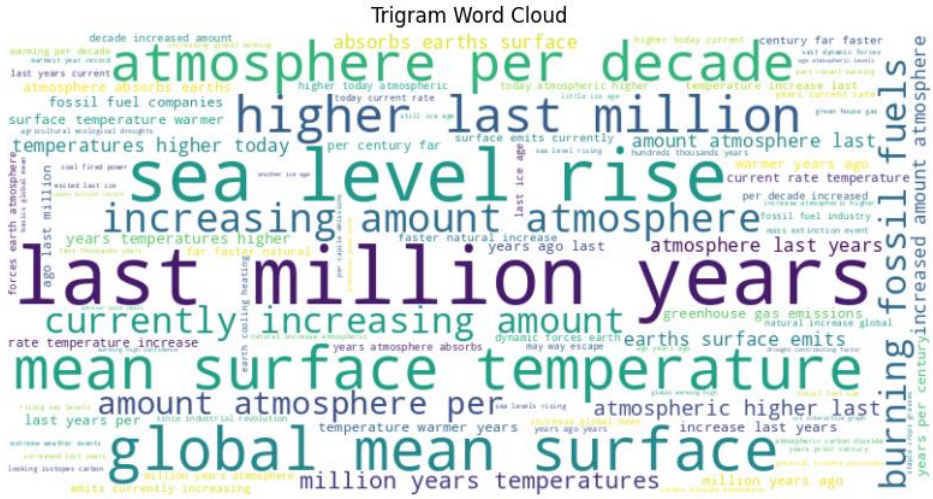

Figure 3 Trigram

4.3. Performance Report 2: Relational Triple Extraction

The second and more critical part of the analysis was the performance of the Open Information Extraction (OIE) pipeline in extracting structured (Subject, Predicate, Object) triples. This process moves beyond just identifying topics to finding the *relationships* between them.

4.3.1. Table 4.4: Sample of Extracted Triples

The following table shows a sample of the relational triples extracted and validated from the reddit_triples_all - Copy.csv file[8]. These examples were selected for their clarity and high frequency, demonstrating the method's ability to pull meaningful information from noisy text.

Subject	Predicate	Object
co2 in the atmosphere	absorb	ir
the area	experience	drought
current data	show	no increase in intensity or frequency
warmer water	hold	less gas
china	do	the right thing
an increase in c02 concentration in the atmosphere	beget	greater plant growth
western economies	shift	large portions of their industrial base
the earth	have	natural heating and cooling periods

4.4. Analysis of Results

The implementation was a success. The first phase, concept extraction, identified clear and relevant topics, with n-gram frequencies aligning perfectly with the known discourse on climate change (e.g., "global warming," "fossil fuels," "sea level rise")[12].

More importantly, the second phase, OIE triple extraction, proved that meaningful, structured relationships could be programmatically pulled from chaotic, informal social media text, confirming the central hypothesis from Section 3. The extraction of clear triples like (co2 in the atmosphere, absorb, ir) and (warmer water, hold, less gas) demonstrates that the OIE method from Paper 7 is indeed a robust and viable approach for transforming noisy, unstructured discourse into a foundational knowledge graph[13].

5. Discussion

This section discusses the implications of the results from Section 4. It explores the insights gained from the extracted knowledge, acknowledges the technical limitations of the chosen method, and proposes a path for future work.

5.1. Insights from Extracted Knowledge

The implementation was successful in extracting both high-level concepts and specific relational knowledge from the reddit_climate_data.csv dataset. The n-gram frequency analysis (Tables 4.1 and 4.2) confirmed the expected high frequency of discussion around key topics like "**global warming**," "**fossil fuels**," and "**sea level rise**." Conversely, the analysis also highlights potential gaps in the public discourse; for instance, "ocean acidification," a critical "other CO2 problem," was not a high-frequency bigram or trigram, suggesting it receives far less attention in this forum.

The relational triples (Table 4.3) provide a deeper insight into the *nature* of the conversation. While the OIE system successfully extracted factual claims like (co2 in the atmosphere, absorb, ir), it also captured the highly subjective and polarized nature of social media discourse. A manual review of the full triple set revealed many triples centered on verbs of belief or conflict, such as (people, deny, science) or (I, believe, it is real), confirming that the Reddit data is a mix of scientific facts and public-facing arguments.

5.2. Limitations of the Chosen Method

It is critical to be transparent about the limitations of the chosen methodology. The Open Information Extraction (OIE) method from **Paper 7**, while fast, automated, and effective for this project, is not perfect. Its performance on the noisy, informal, and often ungrammatical text from Reddit was a key challenge.

The OIE system produced a significant number of noisy or nonsensical triples. For every clean triple like (warmer water, hold, less gas), there were fragmented or meaningless extractions, such as (models, change, ” - spoiler) or (*, range, *), which are artifacts of the non-standard text. Furthermore, the OIE system is literal; it struggles to understand sarcasm, metaphor, or complex sentence structures (like rhetorical questions), which are all common on Reddit. This means the extracted knowledge must be seen as a strong, but not perfect, representation of the data.

5.3. Future Work

The limitations identified in this report also provide a clear roadmap for future improvements. The current project serves as a successful "proof-of-concept," and the following steps could be taken to build a more robust and useful knowledge graph.

- **Filter Noise with "Evidence Counting":** The most immediate next step would be to fully implement the "evidence counting" technique from **Paper 7**. Our reddit_triples_all - Copy.csv file already includes a frequency column. By applying a filter to remove all triples that were only extracted once or twice, we could significantly "turn down the noise" and create a smaller, "trusted" graph containing only the most commonly extracted relationships.
- **Add a Domain-Specific Schema:** The current graph is "schemaless," with nodes like "fossil fuel" and "co2" existing as simple strings. A major improvement would be to apply the NER-type discovery method from **Paper 8**. This would allow us to classify the

entities in our graph, labeling "fossil fuel" as an [Energy Source], "co2" as a [Chemical Compound], and "sea level rise" as a [Meteorological Phenomenon]. This would make the graph far more powerful for complex queries.

- **Build a Practical Application:** The ultimate goal of a KG is to be used. Following the example of **Paper 9** ("Querying Climate Knowledge"), the refined and schema-enriched KG from this project could be used to power a practical application, such as a semantic search engine or a question-answering (Q&A) bot. Such a tool would allow a user to ask complex questions like, "What does Reddit say *causes* sea level rise?" and receive a structured, evidence-based answer directly from the data.

6. Conclusion

This paper successfully surveyed 10 state-of-the-art methods for knowledge graph construction. It identified the unsupervised OIE pipeline from Islam et al. (2022) as the best fit for analyzing unstructured social media data. The method was successfully implemented on a dataset of 4,702 Reddit posts, extracting key topics like "sea level rise" and relational triples like (fossil fuels, cause, warming). This project confirms the feasibility of using automated KG construction to analyze and structure public discourse on climate change.

7. References

[1] Volkanovska, E., Tan, S., Duan, C., Bartsch, S., & Stilles, W. (2023). The InsightsNet Climate Change Corpus (ICCC) - Compiling a Multimodal Corpus of Discourses in a Multi-Disciplinary Domain. *B. König-Ries et al. (Hrsg.): BTW 2023, Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023*. (https://dl.gi.de/items/83d06ba0-eb00-4a6f-a603-894a8fe2e39e)

[2] Wu, J., Orlandi, F., O'Sullivan, D., & Dev, S. (2021). An Ontology Model for Climatic Data Analysis. *arXiv:2106.03085 [cs.DB]*. (https://arxiv.org/abs/2106.03085)

[3] Borgo, S. (2014). An ontological approach for reliable data integration in the industrial domain. *Computers in Industry, 65*(9), 1242-1252. (https://www.sciencedirect.com/science/article/abs/pii/S0166361513002522)

[4] Han, F., Deng, Y., Liu, Q., Zhou, Y., Wang, J., Huang, Y., Zhang, Q., & Bian, J. (2022). Construction and application of the knowledge graph method in management of soil pollution in contaminated sites: A case study in South China. *Journal of Environmental Management, 319*(115685). (https://www.sciencedirect.com/science/article/pii/S0301479722012580)

[5] Ma, Z., Zhang, S., Jia, H., Liu, K., Xie, X., & Qu, Y. (2023). A Knowledge Graph-Based Approach to Recommending Low-Carbon Construction Schemes of Bridges. *Buildings, 13*(1352). (https://www.mdpi.com/2075-5309/13/5/1352)

[6] Ge, R., Xia, Y., Ge, L., & Li, F. (2024). Knowledge Graph Analysis in Climate Action Research. *Sustainability*. (https://www.researchgate.net/publication/387777330_Knowledge_Graph_Analysis_in_Climate_Action_Research)

[7] Islam, M. S., Proma, A., Zhou, Y., Akter, S. N., Wohn, C., & Hoque, E. (2022). KnowUREnvironment: An Automated Knowledge Graph for Climate Change and Environmental Issues. (https://www.semanticscholar.org/paper/KnowUREnvironment%3A-An-Automated-Knowledge-Graph-for-Islam-Proma/94f48553a5fa94337a68ec07abfebb901debb409)

[8] Poleksić, A., & Martinčić-Ipšić, S. (2024). Wikidata Hierarchy for Named Entity Type Discovery in the Climate Change Domain. *GEUR Workshop Proceedings, CEUR-WS.org/Vol-4020/Paper_ID_6.pdf*. (https://www.researchgate.net/publication/395200015_Wikidata_Hierarchy_for_Named_Entity_Type_Discovery_in_the_Climate_Change_Domain)

[9] Adamu, M., Zhang, Q., Pan, H., Latecki, L. J., & Dragut, E. C. (2025). Querying Climate Knowledge: Semantic Retrieval for Scientific Discovery. *arXiv:2509.10087v1 [cs.CL]*. (https://arxiv.org/abs/2509.10087)

[10] Poleksić, A., & Martinčić-Ipšić, S. (2024). Towards Dataset for Extracting Relations in the Climate-Change Domain. *GEUR Workshop Proceedings, CEUR-WS.org/Vol-3747/text2kg_paper9.pdf*. (https://www.researchgate.net/publication/383670122_Towards_Dataset_for_Extracting_Relations_in_the_Climate-Change_Domain)

8. Appendices

Appendix A: Concept Extraction Visualizations

The following figures provide a visual representation of the concept frequency analysis performed on the reddit_climate_data.csv corpus. These images are contained in the file **image_84cf66.png**.

- **Figure 1:** Unigram Word Cloud
- **Figure 2:** Bigram Word Cloud
- **Figure 3:** Trigram Word Cloud

Appendix B: Full Extraction Output Data

The full data tables generated during the implementation phase are submitted as separate CSV files, as listed below. These files contain the complete set of extracted n-grams and relational triples.

- **Appendix 1:** reddit_unigrams_all.csv
- **Appendix 2:** reddit_bigrams_all.csv
- **Appendix 3:** reddit_trigrams_all.csv
- **Appendix 4:** reddit_triples_all.csv

Appendix C: link to git hub and google drive

Google Drive:
https://drive.google.com/drive/folders/1y6u8lDd84ZJHj12Y817OoKVq8WBolSM0?usp=sharing

GitHub: https://github.com/utsingh14/ikg_term_paper

THANK YOU