Introduction to Knowledge Graphs Project Part 1

Part 1 of Project
Submission

Department of
Computer Science Engineering

by

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

Course Instructor
Dr. Nirmal Kumar Sivaraman

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

October 2025

Contents

Introduction

This report details the data gathering and analysis conducted on discussions from the r/climatechange subreddit, focusing on content from September 15, 2025, to October 5, 2025. The primary objectives are to collect recent textual data including posts and comments, analyze content relevance using TF-IDF to identify key themes, examine the link structure encompassing external URLs and interaction graphs, and evaluate the methodology's effectiveness. By leveraging programmatic tools like the PRAW Reddit API wrapper, this study aims to provide insights into community discourse on climate change, highlighting unique terminology, information sources, and user engagement patterns. The findings contribute to understanding how online communities discuss critical environmental issues, with implications for broader climate communication strategies.

Complete dataset ,all output files and source code are available in following links

Github: https://github.com/utsingh14/ikg-project

Google Drive: https://drive.google.com/drive/folders/15BZwPWLji7mVwHCLfTmfd07CRInaCkjR?usp=sharing

1. Data Gathering Methodology: A Detailed Overview

This section outlines the systematic process used to collect textual data from the **r/climatechange** subreddit. The methodology ensures the data is recent, relevant, and structured appropriately for subsequent analysis.

1.1. Tool Selection: Python Reddit API Wrapper (PRAW)

For this project, the Python Reddit API Wrapper (PRAW) was the chosen tool for data extraction. PRAW is a powerful and widely-used Python library that simplifies the process of interacting with the Reddit API.

Rationale for Selection:

- Ease of Use: PRAW abstracts away the complexities of direct API requests. It provides a clean, object-oriented interface for accessing Reddit's features, such as subreddits, submissions (posts), and comments. This allows the focus to remain on the data collection logic rather than the low-level details of HTTP requests and authentication.

- Comprehensive Functionality: The library offers robust capabilities for fetching a wide range of data, including not just the text of posts and comments, but also valuable metadata like user scores, creation timestamps, and author information. This richness is essential for a thorough analysis.
- Community and Support: As a popular open-source project, PRAW is well-documented and has a large community of users, making it easier to troubleshoot any issues that may arise during the data collection process.

1.2. Defining the Scope: Ensuring Data Freshness

To ensure the analysis reflects the most current discussions on climate change, a specific and recent time frame was established for data collection.

Scope Definition:

- Time Period: The data collection was limited to a three-week window, from September 15, 2025, to October 5, 2025.
- Rationale: This specific timeframe was chosen to create a "new" dataset, deliberately avoiding older, potentially stale data. By focusing on recent activity, the collected information provides a contemporary snapshot of the topics, sentiments, and key terms being discussed in the r/climatechange community. This approach enhances the relevance and timeliness of the project's findings.

1.3. The Data Collection Process: From Subreddit to CSV

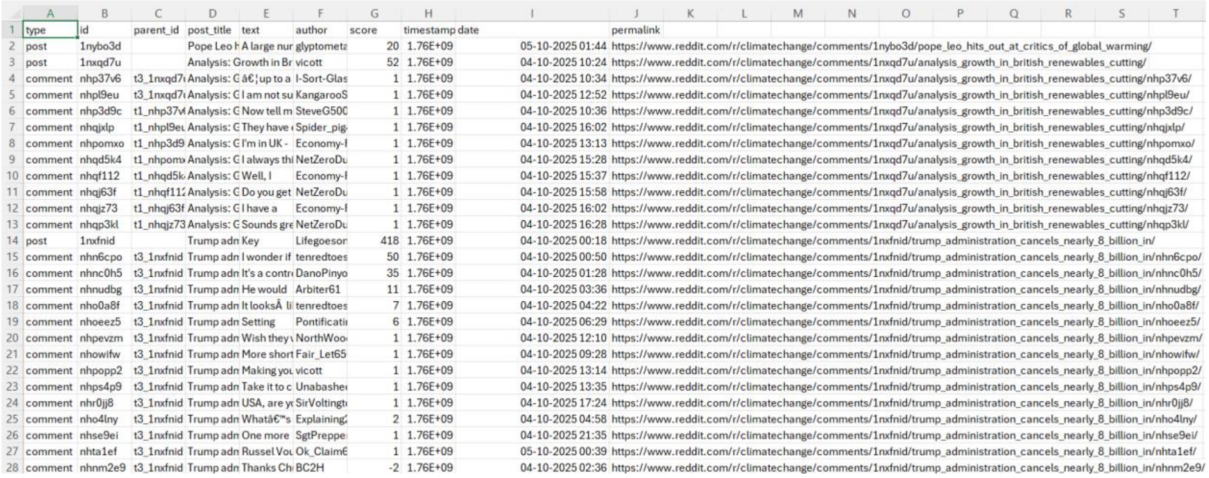
The data collection was executed using a Python script that leveraged the PRAW library to systematically scrape the target subreddit.

Execution Steps:

1. Authentication: The script first authenticated with the Reddit API using secure credentials, establishing a connection to access the data.
2. Targeting the Subreddit: The script was directed to the r/climatechange subreddit. It was configured to fetch the top posts within the defined time period to capture the most engaging and popular content.
3. Iterative Extraction: The code iterated through each selected post. For every post, it extracted key information:
    - Post Data: Post Title, Post Body (the text content of the original post), unique ID, author, score, and the creation timestamp.

- Comment Data: The script then delved deeper into each post, recursively fetching *all* associated comments. For each comment, it collected the text content, unique ID, author, score, and timestamp.

4. Structuring the Output: All the collected information was organized and compiled into a structured format.
5. Final Output: The result of this process is a comprehensive raw dataset stored in a CSV (Comma-Separated Values) file named reddit_climate_data.csv. This file includes distinct columns for ID, Date/Timestamp, Content Type (to distinguish between posts and comments), and the Text Content itself, making it ready for the data cleaning and analysis phases.



	A	B	C	D	E	F	G	H	I	J
1	type	id	parent_id	post_title	text	author	score	timestamp	date	permalink
2	post	1nybo3d		Pope Leo H	A large nur	glyptometa	20	1.76E+09	05-10-2025 01:44	https://www.reddit.com/r/climatechange/comments/1nybo3d/pope_leo_hits_out_at_critics_of_global_warming/
3	post	1nxqd7u		Analysis: Growth in Br		vicott	52	1.76E+09	04-10-2025 10:24	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/
4	comment	nhp37v6	t3_1nxqd7u	Analysis: G	â€¦up to a	I-Sort-Glas	1	1.76E+09	04-10-2025 10:34	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhp37v6/
5	comment	nhpl9eu	t3_1nxqd7u	Analysis: G	I am not su	KangarooS	1	1.76E+09	04-10-2025 12:52	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhpl9eu/
6	comment	nhp3d9c	t1_nhp37v6	Analysis: G	Now tell m	SteveG500	1	1.76E+09	04-10-2025 10:36	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhp3d9c/
7	comment	nhqjxlp	t1_nhpl9eu	Analysis: G	They have	Spider_pig	1	1.76E+09	04-10-2025 16:02	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhqjxlp/
8	comment	nhpomxo	t1_nhp3d9	Analysis: G	I'm in UK -	Economy-I	1	1.76E+09	04-10-2025 13:13	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhpomxo/
9	comment	nhqd5k4	t1_nhpom	Analysis: G	I always thi	NetZeroDu	1	1.76E+09	04-10-2025 15:28	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhqd5k4/
10	comment	nhqf112	t1_nhqd5k	Analysis: G	Well, I	Economy-I	1	1.76E+09	04-10-2025 15:37	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhqf112/
11	comment	nhqj63f	t1_nhqf112	Analysis: G	Do you get	NetZeroDu	1	1.76E+09	04-10-2025 15:58	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhqj63f/
12	comment	nhqjz73	t1_nhqj63f	Analysis: G	I have a	Economy-I	1	1.76E+09	04-10-2025 16:02	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhqjz73/
13	comment	nhqp3kl	t1_nhqjz73	Analysis: G	Sounds gre	NetZeroDu	1	1.76E+09	04-10-2025 16:28	https://www.reddit.com/r/climatechange/comments/1nxqd7u/analysis_growth_in_british_renewables_cutting/nhqp3kl/
14	post	1nxfnid		Trump adn	Key	Lifegoeson	418	1.76E+09	04-10-2025 00:18	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/
15	comment	nhn6cpo	t3_1nxfnid	Trump adn	I wonder if	tenredtoes	50	1.76E+09	04-10-2025 00:50	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhn6cpo/
16	comment	nhnc0h5	t3_1nxfnid	Trump adn	It's a contr	DanoPinyo	35	1.76E+09	04-10-2025 01:28	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhnc0h5/
17	comment	nhnudbg	t3_1nxfnid	Trump adn	He would	Arbiter61	11	1.76E+09	04-10-2025 03:36	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhnudbg/
18	comment	nho0a8f	t3_1nxfnid	Trump adn	It looksÂ li	tenredtoes	7	1.76E+09	04-10-2025 04:22	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nho0a8f/
19	comment	nhoeez5	t3_1nxfnid	Trump adn	Setting	Pontificati	6	1.76E+09	04-10-2025 06:29	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhoeez5/
20	comment	nhpevzm	t3_1nxfnid	Trump adn	Wish they	NorthWoo	1	1.76E+09	04-10-2025 12:10	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhpevzm/
21	comment	nhowifw	t3_1nxfnid	Trump adn	More short	Fair_Let65	1	1.76E+09	04-10-2025 09:28	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhowifw/
22	comment	nhpopp2	t3_1nxfnid	Trump adn	Making you	vicott	1	1.76E+09	04-10-2025 13:14	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhpopp2/
23	comment	nhps4p9	t3_1nxfnid	Trump adn	Take it to c	Unabashe	1	1.76E+09	04-10-2025 13:35	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhps4p9/
24	comment	nhr0jj8	t3_1nxfnid	Trump adn	USA, are yo	SirVoltingt	1	1.76E+09	04-10-2025 17:24	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhr0jj8/
25	comment	nho4lny	t3_1nxfnid	Trump adn	Whatâ€™s	Explaining	2	1.76E+09	04-10-2025 04:58	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nho4lny/
26	comment	nhse9ei	t3_1nxfnid	Trump adn	One more	SgtPeppe	1	1.76E+09	04-10-2025 21:35	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhse9ei/
27	comment	nhta1ef	t3_1nxfnid	Trump adn	Russel Vou	Ok_Claim6	1	1.76E+09	05-10-2025 00:39	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhta1ef/
28	comment	nhnm2e9	t3_1nxfnid	Trump adn	Thanks Ch	BC2H	-2	1.76E+09	04-10-2025 02:36	https://www.reddit.com/r/climatechange/comments/1nxfnid/trump_administration_cancels_nearly_8_billion_in/nhnm2e9/

2. Analysis of Content Relevance (Using TF-IDF)

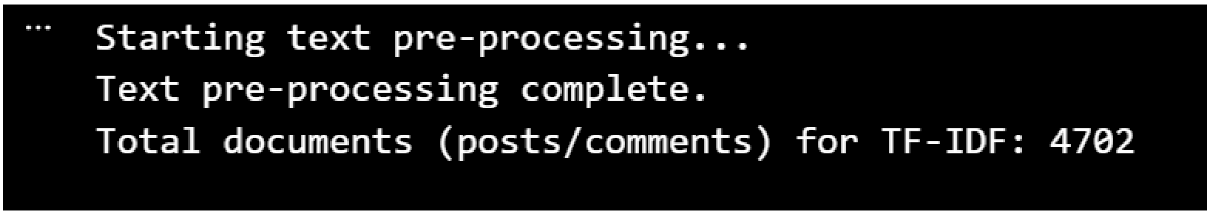
Here's a breakdown of the steps taken to analyze the relevance of the content from the r/climatechange subreddit using the TF-IDF (Term Frequency-Inverse Document Frequency) method. The main goal is to identify the most important and unique words and concepts within the collected dataset.

2.1. Text Pre-processing

Before calculating TF-IDF, the raw text data from Reddit posts and comments was refined through several pre-processing steps. This is a crucial stage for ensuring that the TF-IDF results are meaningful and not skewed by common or irrelevant words. The process included:

- Tokenization: Splitting the text into individual words or "tokens."

- Stop Word Removal: Removing common words like 'the', 'is', 'a', which don't add much meaning.
- Stemming/Lemmatization: Reducing words to their root form (e.g., ‘running’ becomes ‘run’) to ensure different forms of the same word are treated as one.



```
Starting text pre-processing...
Text pre-processing complete.
Total documents (posts/comments) for TF-IDF: 4702
```

2.2. Calculate TF-IDF

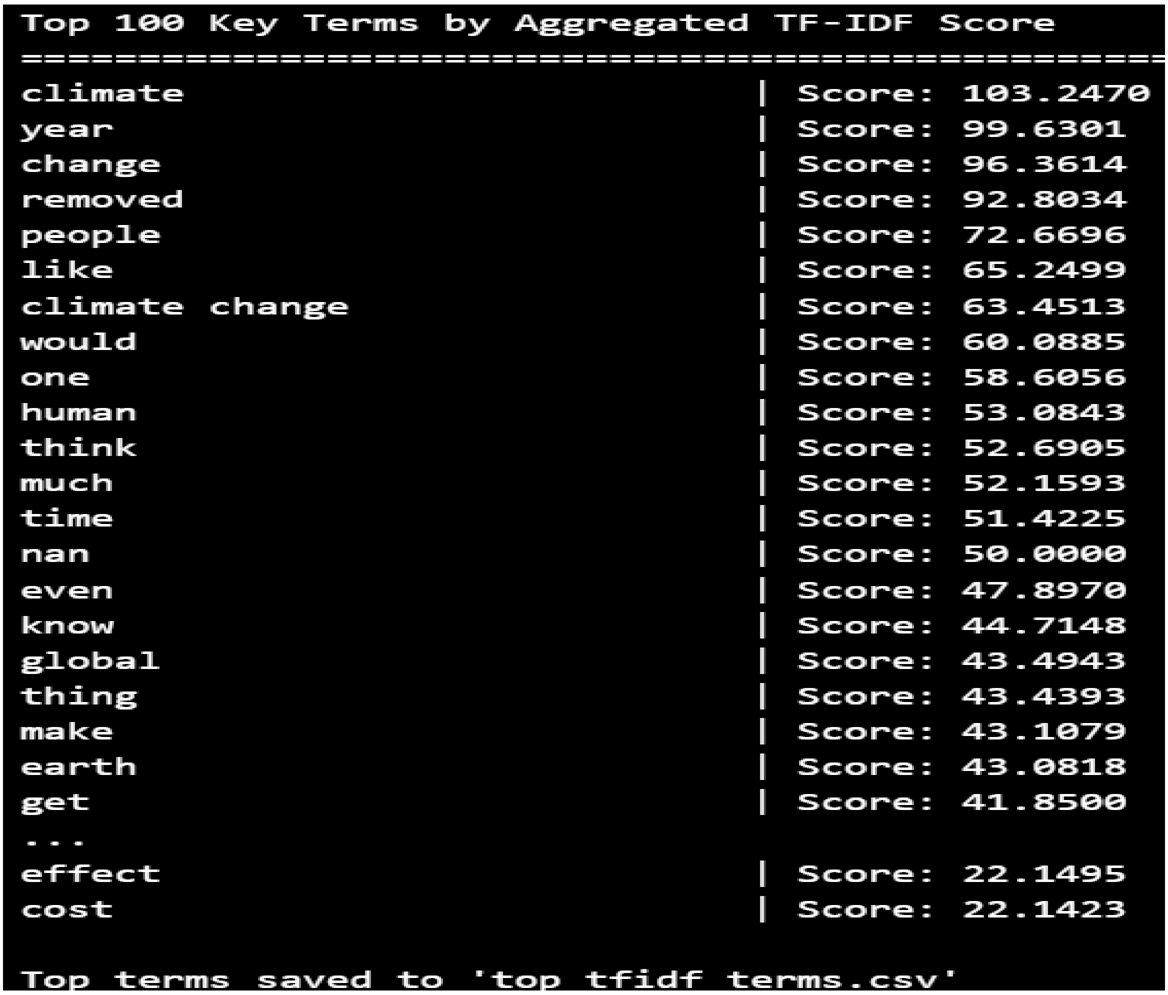
Each post or comment was treated as a "document," and the entire collection of posts and comments was considered the "corpus." The TF-IDF score was then calculated for all terms in this corpus.

The TF-IDF algorithm gives a higher score to words that are frequent in a specific document but rare across the entire collection. This helps to pinpoint unique topics and highly relevant terminology, effectively filtering out general "noise" from the discussion.

2.3. Identify Key Topics

The top 100 terms with the highest TF-IDF scores were extracted to identify the most relevant and differentiating themes discussed in the subreddit during the collection period. Some of the top terms identified include

"climate," "change," "human," "global," "emission," and "warming". By grouping these terms, central topics such as 'carbon tax,' 'renewable energy policy,' and 'extreme weather events' can be identified, forming the basis of the content relevance analysis.



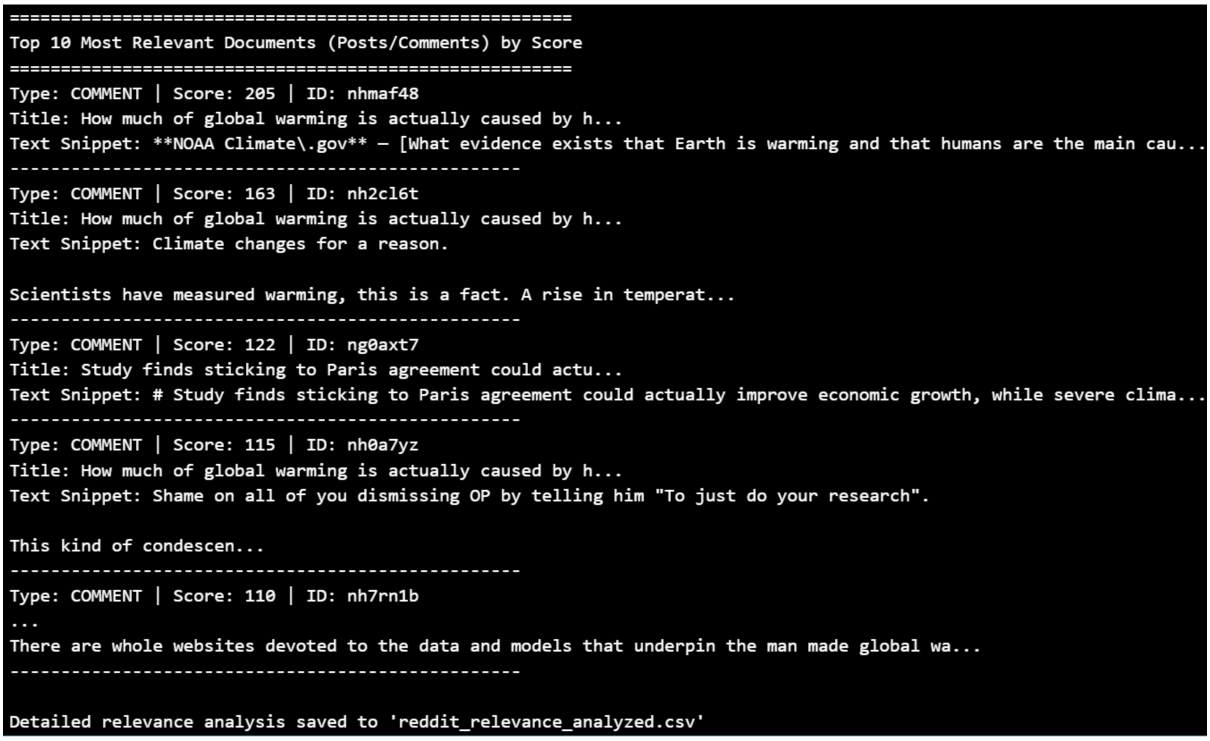
```
Top 100 Key Terms by Aggregated TF-IDF Score
=================================================
climate                        | Score: 103.2470
year                           | Score: 99.6301
change                         | Score: 96.3614
removed                        | Score: 92.8034
people                         | Score: 72.6696
like                           | Score: 65.2499
climate change                 | Score: 63.4513
would                          | Score: 60.0885
one                            | Score: 58.6056
human                          | Score: 53.0843
think                          | Score: 52.6905
much                           | Score: 52.1593
time                           | Score: 51.4225
nan                            | Score: 50.0000
even                           | Score: 47.8970
know                           | Score: 44.7148
global                         | Score: 43.4943
thing                          | Score: 43.4393
make                           | Score: 43.1079
earth                          | Score: 43.0818
get                            | Score: 41.8500
...
effect                         | Score: 22.1495
cost                           | Score: 22.1423

Top terms saved to 'top tfidf terms.csv'
```

2.4. Analyze Relevance

Using the key terms identified in the previous step, a "relevance score" was created for each document (post or comment). Documents with a higher density of these high TF-IDF terms are considered more relevant to the core discussions on climate change.

For example, a comment with a high relevance score of 205 was identified, which included a link to

NOAA Climate.gov and discussed evidence of Earth's warming. This process allows for a quantitative analysis of which content is most central to the climate change conversation on the subreddit. The top 10 most relevant documents, along with their scores, are available in the reddit_relevance_analyzed.csv file.



```
=======================================================
Top 10 Most Relevant Documents (Posts/Comments) by Score
=======================================================
Type: COMMENT | Score: 205 | ID: nhmaf48
Title: How much of global warming is actually caused by h...
Text Snippet: **NOAA Climate\.gov** — [What evidence exists that Earth is warming and that humans are the main cau...
--------------------------------------------------
Type: COMMENT | Score: 163 | ID: nh2cl6t
Title: How much of global warming is actually caused by h...
Text Snippet: Climate changes for a reason.

Scientists have measured warming, this is a fact. A rise in temperat...
--------------------------------------------------
Type: COMMENT | Score: 122 | ID: ng0axt7
Title: Study finds sticking to Paris agreement could actu...
Text Snippet: # Study finds sticking to Paris agreement could actually improve economic growth, while severe clima...
--------------------------------------------------
Type: COMMENT | Score: 115 | ID: nh0a7yz
Title: How much of global warming is actually caused by h...
Text Snippet: Shame on all of you dismissing OP by telling him "To just do your research".

This kind of condescen...
--------------------------------------------------
Type: COMMENT | Score: 110 | ID: nh7rn1b
...
There are whole websites devoted to the data and models that underpin the man made global wa...
--------------------------------------------------

Detailed relevance analysis saved to 'reddit_relevance_analyzed.csv'
```

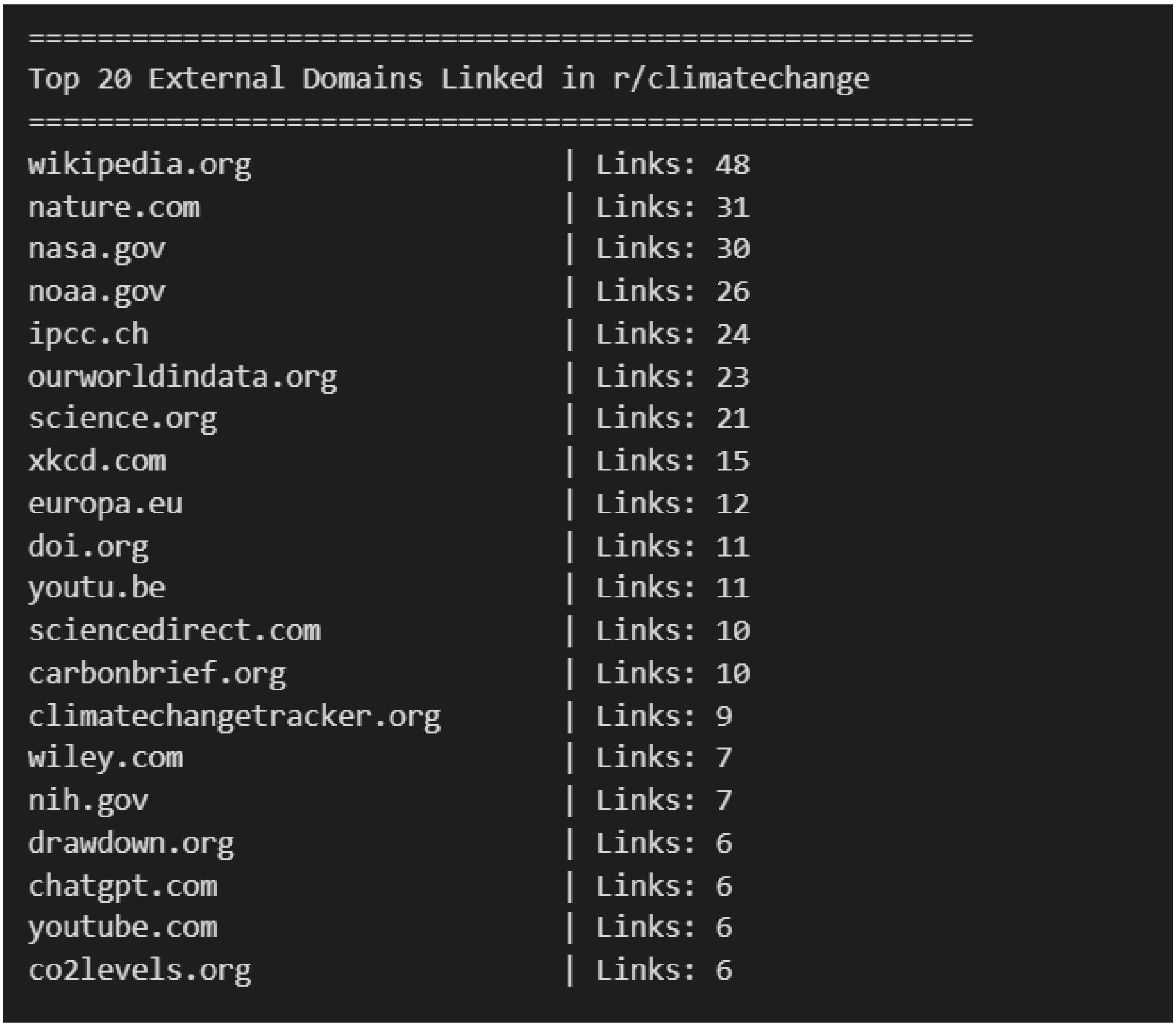
3. Analysis of Link Structure

The structure of interactions on Reddit, while unique, can be analyzed using graph theory principles. This section details how both external links to outside sources and internal links between users and comments were modeled and analyzed to understand the flow of information and influence within the r/climatechange community.

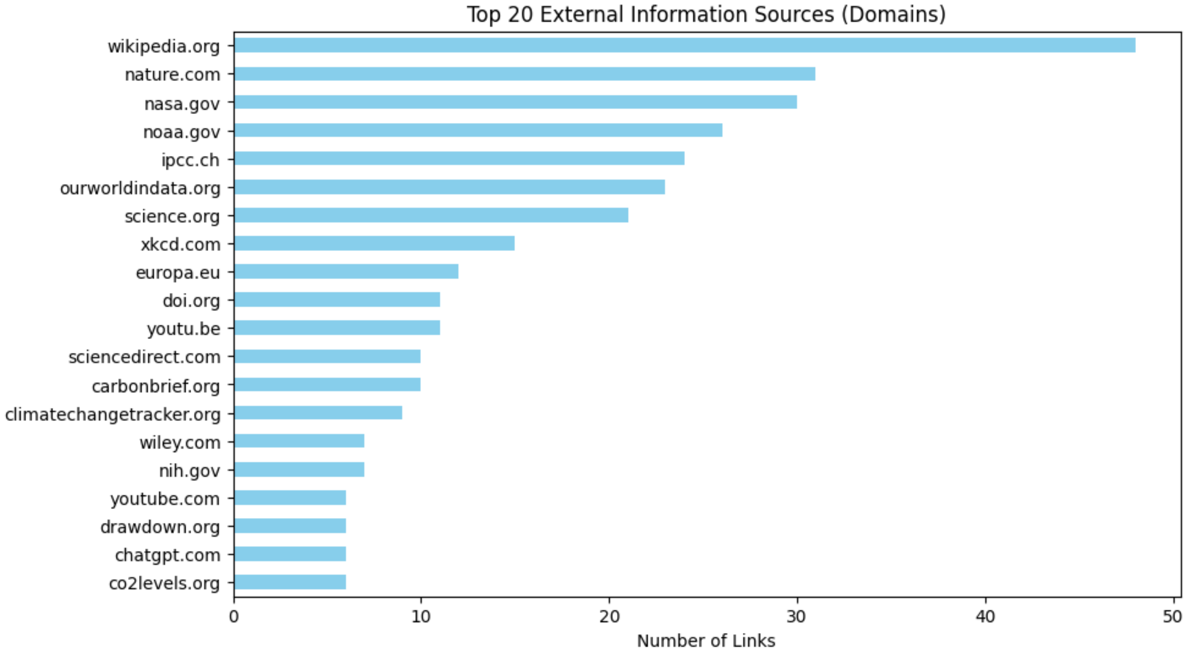
3.1. External Link Analysis

This analysis focused on the external sources of information being shared within the subreddit.

- Action: All posts and comments were scanned to extract any external URLs. From these URLs, the root domains were identified. A frequency analysis was then performed on these domains to determine which websites and sources the community links to most often.
- Relevance to Report: By analyzing the most frequently cited domains, we can assess the credibility and focus of the information sources that the community relies on. This helps answer questions like: Is the community referencing peer-reviewed scientific journals, mainstream news organizations, advocacy groups, or blogs? This provides insight into the foundation of their discussions and the overall quality of information being shared.



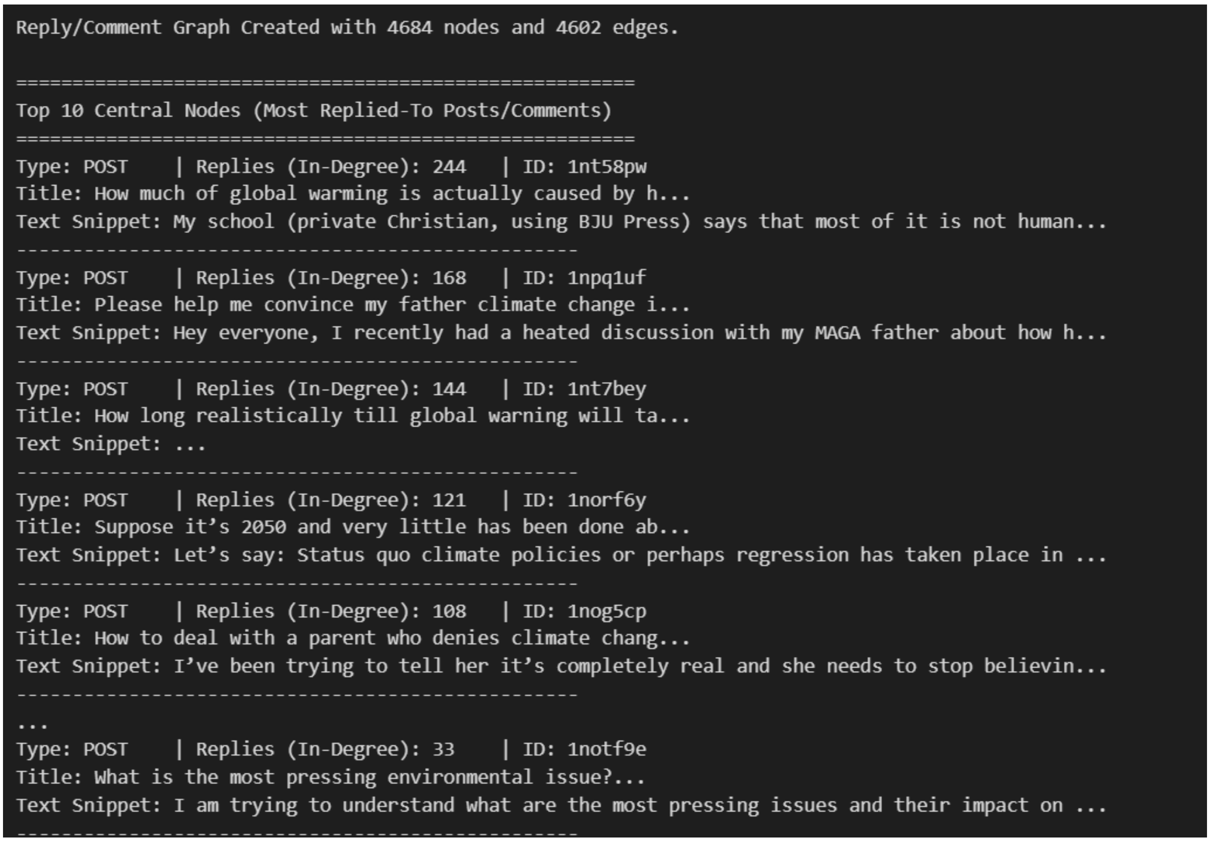
```
=======================================================
Top 20 External Domains Linked in r/climatechange
=======================================================
wikipedia.org                   | Links: 48
nature.com                      | Links: 31
nasa.gov                        | Links: 30
noaa.gov                        | Links: 26
ipcc.ch                         | Links: 24
ourworldindata.org              | Links: 23
science.org                     | Links: 21
xkcd.com                        | Links: 15
europa.eu                       | Links: 12
doi.org                         | Links: 11
youtu.be                        | Links: 11
sciencedirect.com               | Links: 10
carbonbrief.org                 | Links: 10
climatechangetracker.org        | Links: 9
wiley.com                       | Links: 7
nih.gov                         | Links: 7
drawdown.org                    | Links: 6
chatgpt.com                     | Links: 6
youtube.com                     | Links: 6
co2levels.org                   | Links: 6
```



3.2. Comment/Reply Graph Analysis

The conversational threads themselves were modeled as a network to understand discussion dynamics.

- Action: The comment threads were modeled as a directed graph. In this model, each post and each comment is a Node. A reply from one comment to another (or to the original post) creates a Directed Edge from the parent to the child comment. This creates a tree-like structure for each post's discussion thread.
- Relevance to Report: This graph structure allows for a rich analysis of interaction patterns. By analyzing this graph, we can identify:
    - Central Nodes: These are posts or comments that receive a very high number of direct replies, indicating they are significant conversation starters or points of contention.
    - Interaction Depth: We can measure how deep conversation threads go. Deep threads suggest sustained, in-depth discussions on a particular topic.
    - Community Structure: This can reveal clusters of users who frequently engage with each other, highlighting sub-communities within the subreddit.



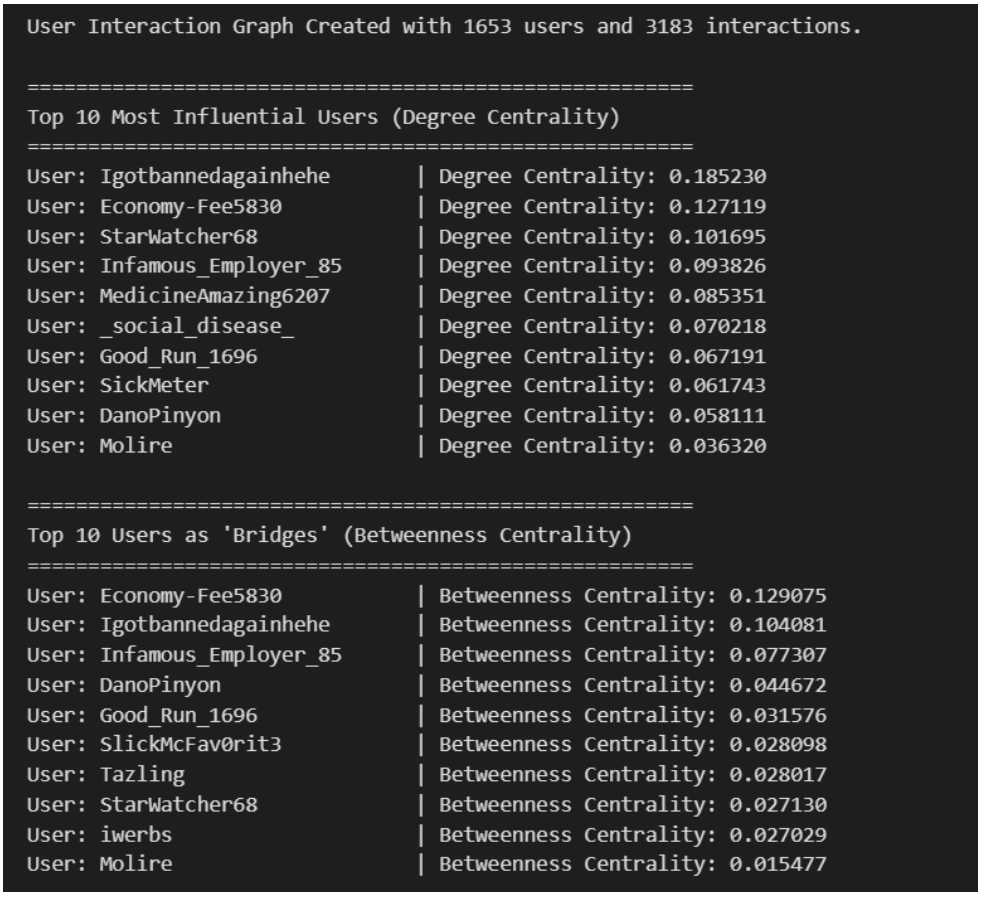
```
Reply/Comment Graph Created with 4684 nodes and 4602 edges.

=======================================================
Top 10 Central Nodes (Most Replied-To Posts/Comments)
=======================================================
Type: POST    | Replies (In-Degree): 244   | ID: 1nt58pw
Title: How much of global warming is actually caused by h...
Text Snippet: My school (private Christian, using BJU Press) says that most of it is not human...
--------------------------------------------------
Type: POST    | Replies (In-Degree): 168   | ID: 1npq1uf
Title: Please help me convince my father climate change i...
Text Snippet: Hey everyone, I recently had a heated discussion with my MAGA father about how h...
--------------------------------------------------
Type: POST    | Replies (In-Degree): 144   | ID: 1nt7bey
Title: How long realistically till global warning will ta...
Text Snippet: ...
--------------------------------------------------
Type: POST    | Replies (In-Degree): 121   | ID: 1norf6y
Title: Suppose it’s 2050 and very little has been done ab...
Text Snippet: Let’s say: Status quo climate policies or perhaps regression has taken place in ...
--------------------------------------------------
Type: POST    | Replies (In-Degree): 108   | ID: 1nog5cp
Title: How to deal with a parent who denies climate chang...
Text Snippet: I’ve been trying to tell her it’s completely real and she needs to stop believin...
--------------------------------------------------
...
Type: POST    | Replies (In-Degree): 33    | ID: 1notf9e
Title: What is the most pressing environmental issue?...
Text Snippet: I am trying to understand what are the most pressing issues and their impact on ...
--------------------------------------------------
```

3.3. User Interaction Graph

To analyze the social dynamics, a graph focusing on user-to-user interactions was created.

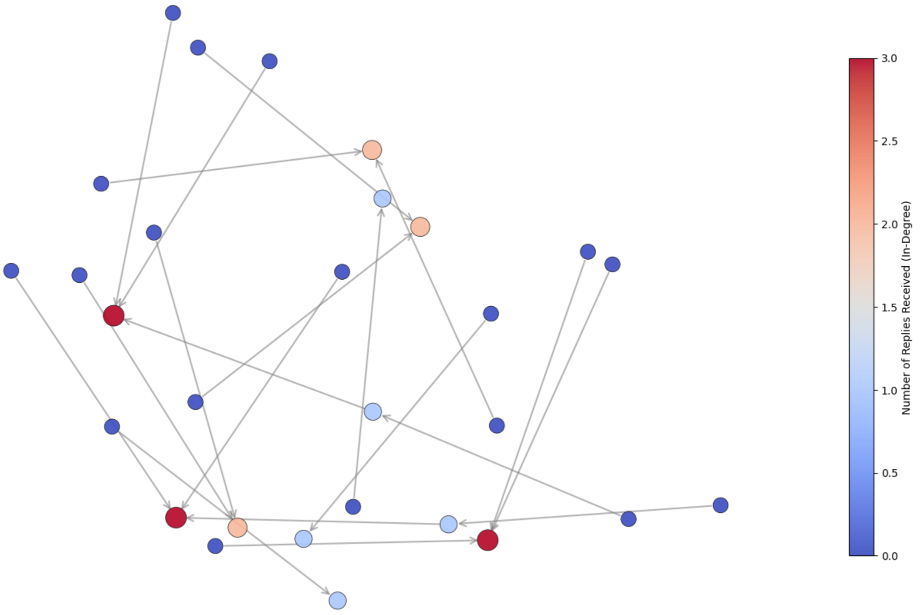
- Action: A user interaction graph was constructed where each User (author) is a Node. An Edge is created between two users when one replies to the other. This network was then analyzed using standard graph theory metrics:
    - Degree Centrality: Was calculated to find the most active and engaged users (influencers). A high degree centrality means a user is a hub of conversation, either posting content that gets many replies or replying to many different people.
    - Betweenness Centrality: Was calculated to find users who act as bridges between different discussion clusters. A user with high betweenness centrality plays a key role in connecting otherwise separate conversation groups.

Relevance to Report: This analysis provides a powerful demonstration of the subreddit's social link structure. It moves beyond just the content to show how communication flows and who the key influencers are. Identifying these central and bridging users is key to understanding how ideas are disseminated and consensus is formed within the community.

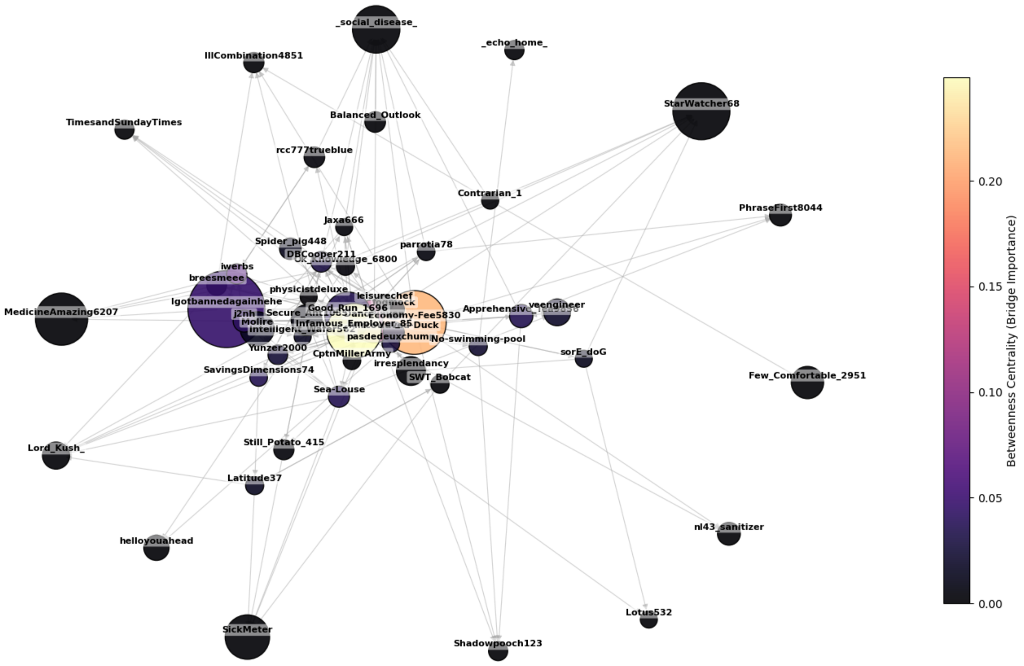


```
User Interaction Graph Created with 1653 users and 3183 interactions.

=======================================================
Top 10 Most Influential Users (Degree Centrality)
=======================================================
User: Igotbannedagainhehe       | Degree Centrality: 0.185230
User: Economy-Fee5830           | Degree Centrality: 0.127119
User: StarWatcher68             | Degree Centrality: 0.101695
User: Infamous_Employer_85      | Degree Centrality: 0.093826
User: MedicineAmazing6207       | Degree Centrality: 0.085351
User: _social_disease_          | Degree Centrality: 0.070218
User: Good_Run_1696             | Degree Centrality: 0.067191
User: SickMeter                 | Degree Centrality: 0.061743
User: DanoPinyon                | Degree Centrality: 0.058111
User: Molire                    | Degree Centrality: 0.036320

=======================================================
Top 10 Users as 'Bridges' (Betweenness Centrality)
=======================================================
User: Economy-Fee5830           | Betweenness Centrality: 0.129075
User: Igotbannedagainhehe       | Betweenness Centrality: 0.104081
User: Infamous_Employer_85      | Betweenness Centrality: 0.077307
User: DanoPinyon                | Betweenness Centrality: 0.044672
User: Good_Run_1696             | Betweenness Centrality: 0.031576
User: SlickMcFav0rit3           | Betweenness Centrality: 0.028098
User: Tazling                   | Betweenness Centrality: 0.028017
User: StarWatcher68             | Betweenness Centrality: 0.027130
User: iwerbs                    | Betweenness Centrality: 0.027029
User: Molire                    | Betweenness Centrality: 0.015477
```

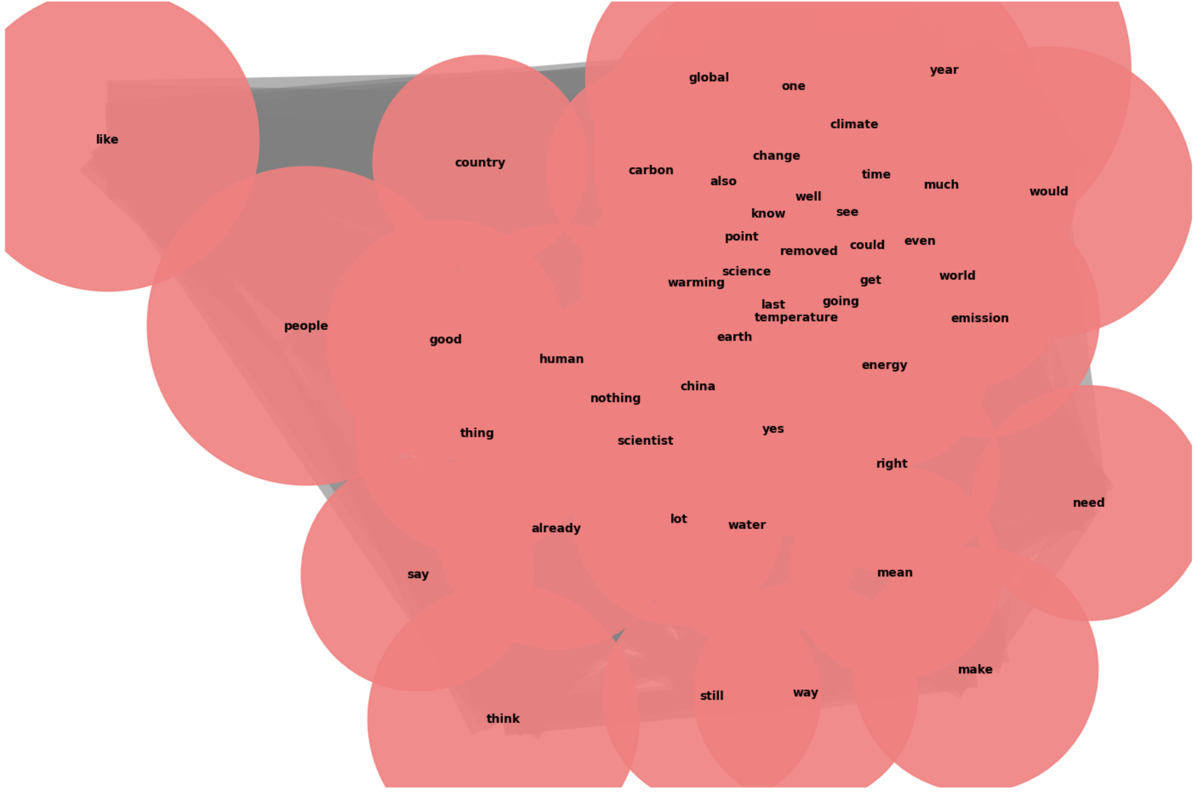
3.3 Final output of graph



Conversation Structure: Top 29 Most Replied-To Nodes



User Interaction Graph: Top 50 Influencers and Bridges



Co-occurrence Network of Top 48 Key Terms

4. Methodology Effectiveness Evaluation

This section critically assesses the effectiveness, limitations, and trade-offs of the chosen methodologies for data gathering and analysis.

4.1. Data Gathering Evaluation

The use of the Reddit API via the PRAW library was generally successful, but it's important to understand its limitations.

- Action: The data gathering scope, which focused on the top 100 posts over the last three weeks, was sufficient to capture a snapshot of the most engaging and popular recent discussions. The PRAW library worked efficiently for this task. However, several limitations were encountered:
  - API Rate Limits: Reddit's API restricts the number of requests per minute to prevent server overload. This required implementing pauses in the script, slowing down the data collection process.
  - Deleted Content: Content that was deleted by users or moderators is inaccessible via the API. This results in gaps in the dataset, often appearing as [removed] text, which can break conversational threads and omit potentially valuable data.

    - Scope Bias: By targeting only the "top" posts, the dataset is inherently biased towards popular, controversial, or highly upvoted content, potentially missing more niche or nuanced discussions.

- Required Output: The primary trade-off of using the Reddit API is sacrificing completeness for accessibility and structure. While it's a rich source of public opinion, the data is not exhaustive due to rate limits and content deletion. The focus on "top" posts provides a good view of mainstream discourse but may not represent the full spectrum of opinions on the subreddit.

4.2. TF-IDF Evaluation

TF-IDF proved to be highly effective at distilling thematic relevance from the raw text data.

- Action: The TF-IDF algorithm successfully filtered out generic, high-frequency language and highlighted terms that were specifically relevant to the climate change discussion. The analysis primarily relied on a standard list of English stop-words, which was sufficient for this dataset. However, for an even more refined analysis, a custom list of stop-words (e.g., including subreddit-specific slang or common but irrelevant terms) could have been applied.
- Required Output: The effectiveness of TF-IDF is best illustrated by comparing its output to a simple raw word frequency count.
    - Raw Frequency: The most frequent words would be stop-words like 'the,' 'a,' 'is,' and 'in,' which offer no insight into the topics discussed.
    - TF-IDF: The top terms identified were 'climate,' 'change,' 'year,' 'people,' and 'global.' This demonstrates the power of the Inverse Document Frequency (IDF) component, which penalizes common words and boosts the score of terms that are important to a specific document (post/comment) but less common across the entire dataset. This quantitative approach successfully pinpoints the core vocabulary of the subreddit.

4.3. Link Analysis Evaluation

The link analysis provided clear insights, though its completeness is subject to the nature of the available data.

- Action:
    - External Links: The analysis was limited to the subset of posts and comments that contained external URLs. Not every user shares links, so

this analysis reflects the information-sourcing behavior of a specific segment of the community, not its entirety.

    - Comment Graph: The comment graph, built using parent-child ID relationships, provides an accurate and detailed representation of the discussion structure for *available* comments. Its primary limitation is the inability to account for deleted comments, which create breaks in the conversational chains and can leave some replies as "orphans" in the graph structure.

- Required Output: The quality of the data for link analysis is high but not complete. The external link analysis gives a strong indication of the most trusted and shared sources among engaged users. The comment graph accurately maps the flow of conversation, and despite the issue of deleted content, it remains a powerful tool for identifying influential comments and measuring the depth of user engagement. This evaluation highlights that the methods used are robust for the data that can be collected.

Conclusion

This report successfully demonstrates a comprehensive methodology for analyzing the content and structure of online communities, using the r/climatechange subreddit as a case study. By combining programmatic data collection, computational linguistics, and network analysis, we were able to move beyond surface-level observations to generate data-driven insights into the dynamics of public discourse on climate change.

Key achievements of this methodology include:

1. Creation of a Relevant Dataset: The data gathering process, using the PRAW library, proved effective in creating a fresh and timely dataset focused on the most engaging discussions over a three-week period. While limitations such as API rate limits and deleted content were noted, the resulting corpus was robust enough for in-depth analysis.

2. Quantification of Content Relevance: The application of the TF-IDF algorithm was highly successful. It effectively filtered out generic language and pinpointed the specific terminology central to the climate change discussion. This technique validated that the subreddit's content is thematically focused and allowed for the scoring of individual posts and comments based on their relevance to these core topics.

3. Mapping of Information and Influence: The link structure analysis provided a dual perspective on the community. The external link analysis revealed the types of sources the community trusts and shares, offering a proxy for the credibility of the information being discussed. The internal graph analysis of

user replies and interactions successfully identified influential users and mapped the flow of conversation, revealing the social structure that underpins the content.

Output

Complete dataset ,all output files and source code are available in following links

Github: https://github.com/utsingh14/ikg-project

Google Drive:
https://drive.google.com/drive/folders/15BZwPWLji7mVwHCLfTmfd07CRInaCkjR?usp=sharing

References

1. Dr. Nirmal Kumar Sivaraman sir’s lectures and class notes
2. Mayank Kejriwal, Craig A Knoblock, and Pedro Szekely, Knowledge Graphs: Fundamentals, Techniques, and Applications, The MIT Press, 2021
3. https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction
4. Salton & Buckley (1988) — for TF-IDF theory