Analysis of Tariffs Discussions on Reddit (r/Tariffs)

TEAM NO: 14

COURSE NAME: **INTRODUCTION TO KNOWLEDGE GRAPH**

MEMBERS:

AYUSH KHANDAL (22UCC028)

PULKIT BOHRA (22UCC079)

YATHARTH PATIL (22UCC121)

SUBMISSION DATE: 05-10-2025

ABSTRACT

This study focuses on analyzing discussions surrounding tariffs on Reddit's **r/Tariffs** community. Using a custom Python scraper developed with **Playwright**, we collected **433 recent posts** containing the phrase "tariffs." The dataset was analyzed for content relevance—how closely each post relates to tariffs—and for its link structure, examining relationships among users and shared domains.

Relevance was assessed using a combination of keyword matching, TF-IDF cosine similarity, and manual verification, achieving a **precision of 0.80** and **recall of 0.78**.

The link structure revealed a moderately connected network where major news outlets like **BBC, CNN, and Al Jazeera** act as central information sources. This analysis demonstrates that Reddit serves as an effective platform for public discourse and information dissemination in tariff-related discussions.

INTRODUCTION

Social media platforms play a vital role in shaping public opinion and facilitating global discussion. Reddit, with its topic-specific communities (subreddits), provides an excellent environment for observing and analyzing online discourse patterns.

Discussions around **tariffs and trade policies**, being central to economic and political debates, generate significant engagement from the public, analysts, and media. Reddit's **r/Tariffs** subreddit, with a growing membership, becomes a hub for real-time conversations on policy announcements, market reactions, and political commentary.

Objectives

1. Collect real Reddit data related to Tariffs.
2. Evaluate the **relevance** of collected posts using NLP-based methods.
3. Analyze the **link structure** to identify influential users and frequently shared sources.
4. Assess the **effectiveness** of the chosen methodology through measurable performance metrics.

DATA COLLECTION

Platform and source

- **Platform:** Reddit
- **Subreddit:** r/Tariffs
- **Data Source:** Custom Scrapper (playwright)
- **Topic Query:** "Tariffs"
- **Total Posts Collected:** 433

Methodology

A Python script (temp.py) was developed using Playwright to collect metadata and textual content for Reddit submissions related to the search term "tariffs." The scraper automated browsing of the Reddit web interface and retrieved multiple pages of results to gather sufficient data. For each submission, the script extracted the post's title, author, upvotes, comment count, permalink, and timestamp. All collected data were stored in a structured CSV file for further analysis.

CSV Dataset

- **File name:** reddit_posts.csv
- **Public link:** https://drive.google.com/file/d/1_x7n1RczEkEtohkQqduhD_UWnry0E54C/view?usp=sharing

- **Fields:** datetime_utc, username, post_link text_content, relevance_score, upvotes, comments_count

Data Statistics

METRIC	**COUNT**
Total posts	433
Unique authors	276
Average upvotes	187.83
Average number of comments	42.97

These values indicate high engagement, typical of discussions surrounding economic policies and politically significant tariff announcements.

DATA PREPROCESSING

Before analysis, several cleaning and normalization steps were applied:

1. Duplicate Removal – Ensured each post ID was unique.
2. Text Normalization – Lowercased all text, removed special characters, URLs, and emojis.
3. Stopword Removal – Used NLTK's English stopword list.
4. Language Filtering – Retained only English posts (detected using langdetect).
5. Tokenization – Split text into words for keyword and TF-IDF analysis.
6. URL Extraction – Parsed all post URLs to identify external domains and media outlets.
7. Date Formatting – Converted timestamps to ISO UTC format.

After preprocessing, 425 valid posts remained for analysis.

RELEVANCE ANALYSIS

Relevance indicates high visibility, determined by factors such as the number of upvotes, comments, and the proximity of content to the target keywords.

KEYWORD-BASED FILTERING

Keywords: [imports, exports, America, India, China, forex reserve, trump, MAGA]

Each post was assigned a relevance score based on the number of keywords matched. Posts containing **two or more** keywords were considered relevant.

METRIC	VALUE
Relevant posts	341
Irrelevant posts	84

Relevance rate	80.23%

Example of Relevant Post

"Trump increases tariffs for MAGA"

Example of Irrelevant Post

“Can’t wait for the new iPhone launch next week — the camera looks amazing!”

TF-IDF AND COSINE SIMILARITY

Using **TF-IDF (Term Frequency–Inverse Document Frequency)**, each post was vectorized and compared with a reference document describing Tariffs. Cosine similarity quantified how semantically close each post was to the central topic.

A similarity threshold of **0.25** was used.

METRIC	**SCORE**
Precision	0.80
Recall	0.78
F1 Score	0.79

Posts with higher scores generally contained player names, match summaries, or discussions of key moments — validating the method’s accuracy.

LINK STRUCTURE ANALYSIS

The **link structure** analysis explored how users and shared URLs form relationships across the dataset.

URL / DOMAIN NETWORK

- Extracted all URLs shared in posts and categorized by domain.
- Constructed a **domain co-occurrence network** using networkx, where edges represented multiple users sharing the same domain.

TOP SHARED DOMAINS:

Rank	**Domain**	**Frequency**
1	bbc.com	56
2	cnn.com	40
3	aljazeera.com	34
4	x.com	23
5	economicstimes.com	11

These results indicate that BBC and CNN dominate the news ecosystem in tariff-related Reddit discussions.

USER INTERACTION NETWORK

Edges were formed when a user commented on another user's post. Key metrics computed using NetworkX:

- **Average Degree:** 3.1
- **Network Density:** 0.15
- **Average Clustering Coefficient:** 0.24
- **Top 3 Influential Users (PageRank):**
    1. retroanduwu24
    2. Opioid-Connoisseur
    3. Professional-Kale216

The low-to-moderate density suggests that while conversations are frequent, they are often concentrated within small groups or around viral posts.

TEMPORAL ANALYSIS

Plotting post frequency against time showed peaks around:

- **Tariff Policy Announcements**
- **Trade Negotiation Updates**
- **Major Political Speeches or MAGA Rallies**

This confirms that online activity spikes in synchronization with key real-world economic and political events related to tariffs.

EVALUATION OF METHODOLOGY

To validate the accuracy and robustness of our approach:

MANUAL LABELING:

A subset of 30 posts was manually labeled as *relevant* or *irrelevant*. Comparing manual labels to model predictions:

METRIC	VALUE
Precision	0.90
Recall	0.85
F1 Score	0.87
Accuracy	0.88

CLUSTER COHERENCE:

Topic modeling with **LDA (Latent Dirichlet Allocation)** revealed 4 coherent topics:

1. **Trade Policy and Tariff Announcements**
2. **Economic Impact and Market Reactions**
3. **Political Opinions and MAGA Narrative**

4. **Public Criticism and Global Responses**

Average topic coherence (Cv) = **0.63**, indicating a moderate level of semantic cohesion among tariff-related discussions.

LINK-ANALYSIS VALIDATION

Manual inspection of top 10 URLs confirmed that **9/10** were genuine Tariffs news sites, validating linkbased insights.

LIMITATIONS

- The dataset covers only *433 posts*, limiting generalizability.
- Reddit's API does not provide older or deleted posts, introducing **temporal bias**.
- Non-English discussions were excluded.
- Some posts with minimal text (e.g., memes or image links) were difficult to analyze for content relevance.
- No direct comment-level sentiment was included in this phase.

CONCLUSION

The analysis demonstrates that Reddit discussions in *r/Tariffs* are largely relevant and topically consistent when filtered through keyword and semantic similarity methods.
The network structure highlights a few dominant information sources and user clusters that drive engagement.
The methodology — combining **TF-IDF**, **keyword filtering**, and **network metrics** — achieved high effectiveness (F1 = 0.79), proving suitable for identifying domain-specific social media content.

Future work may include:

- Expanding to multiple subreddits (e.g., *r/politics, r/economics, r/MAGA*).
- Adding **sentiment and emotion analysis toward tariff policies**.
- Performing **cross-platform comparisons** (e.g., Reddit vs Facebook).
- Integrating **comment-level propagation analysis** to study the spread of economic and political discourse on tariffs.