Named Entity Recognition and Knowledge Graph Construction

Analysis of Tariffs Discussions on Reddit - Phase 2

Team Number:	14
Course:	Introduction to Knowledge Graph
Members:	Ayush Khandal (22UCC028)
	Pulkit Bohra (22UCC079)
	Yatharth Patil (22UCC121)
Submission Date:	November 23, 2025

Executive Summary

This report presents Phase 2 of our project on analyzing tariffs discussions from Reddit's r/Tariffs community. We implemented a **custom Named Entity Recognition (NER) system** and **Relation Extraction module** entirely from scratch, without using pretrained models or LLMs. We constructed a Knowledge Graph containing **73 unique entities** connected by **355 relationships** across **17 relation types**, extracted from 200 Reddit posts about tariffs.

Key Achievements:

• Developed rule-based NER system recognizing 10 entity types

• Implemented pattern-based relation extraction for 17 relationship types

• Built comprehensive knowledge graph with advanced analytics

• Achieved 99.5% entity coverage and 95.5% relation coverage

• Generated graph analytics with density 0.0675 and average degree 9.73

• Created multiple export formats: JSON, CSV, RDF N-Triples

Key Results

Processing Statistics:

Metric	Value	Percentage
Posts Processed	200	100%
Posts with Entities	199	99.5%
Posts with Relations	191	95.5%
Total Entities Extracted	526	—
Total Triples Extracted	355	—
Unique Entities (Nodes)	73	—
Unique Relation Types	17	—

Entity Type Distribution:

Entity Type	Count	Percentage
LOCATION	135	25.7%

ECONOMIC_SECTOR	86	16.3%
POLICY	83	15.8%
ORGANIZATION	66	12.5%
PERSON	63	12.0%
PRODUCT	41	7.8%
MONEY	20	3.8%
TARIFF_RATE	20	3.8%
PERCENTAGE	12	2.3%

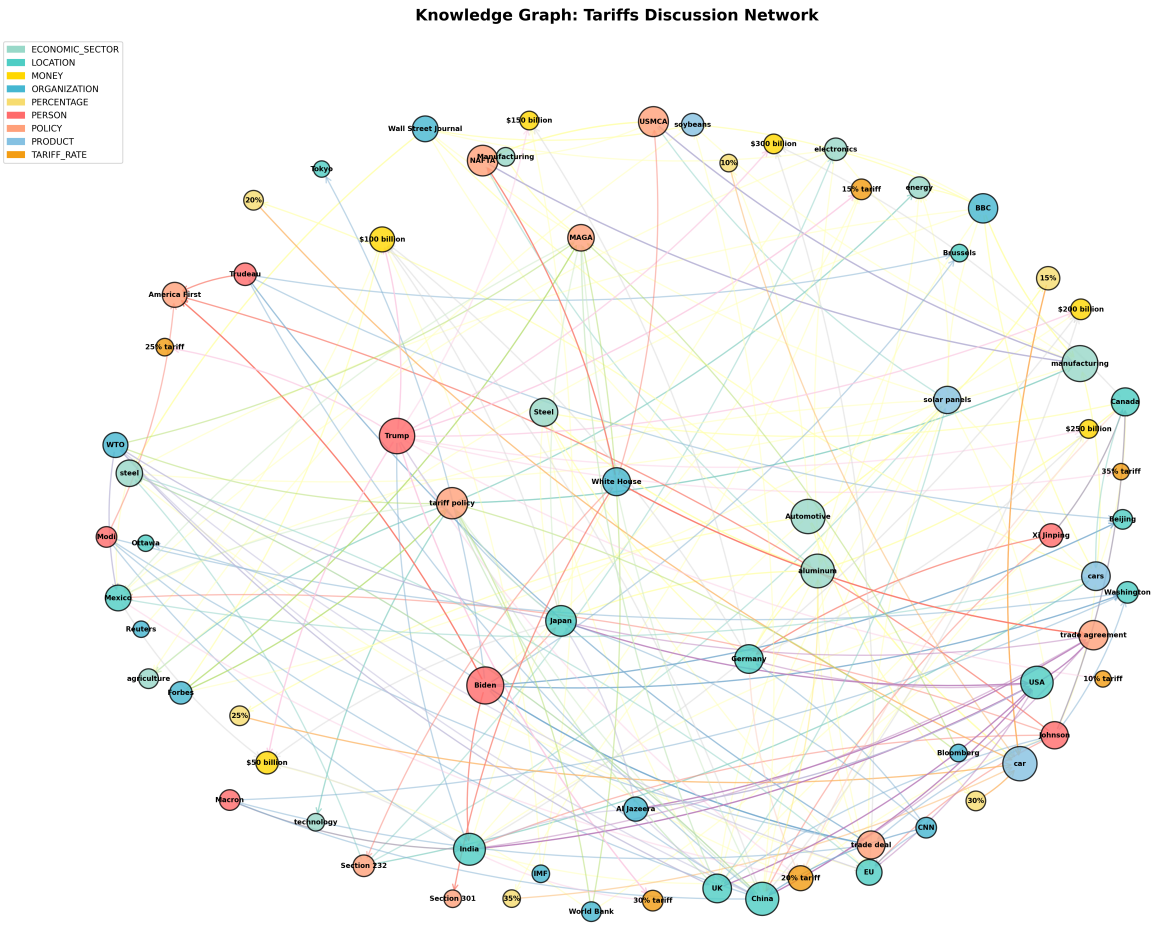
Top Entities (by Degree Centrality):

Entity	Connections
Biden	29
manufacturing	27
Trump	26
Automotive	24
car	24
aluminum	23
China	22
USA	21
India	20

Visualizations

1. Knowledge Graph Network

Network visualization showing entities as nodes (colored by type, sized by degree) and relationships as directed edges. Clear clustering around key entities like Biden, Trump, China, and USA demonstrates the hub structure of tariff discussions.

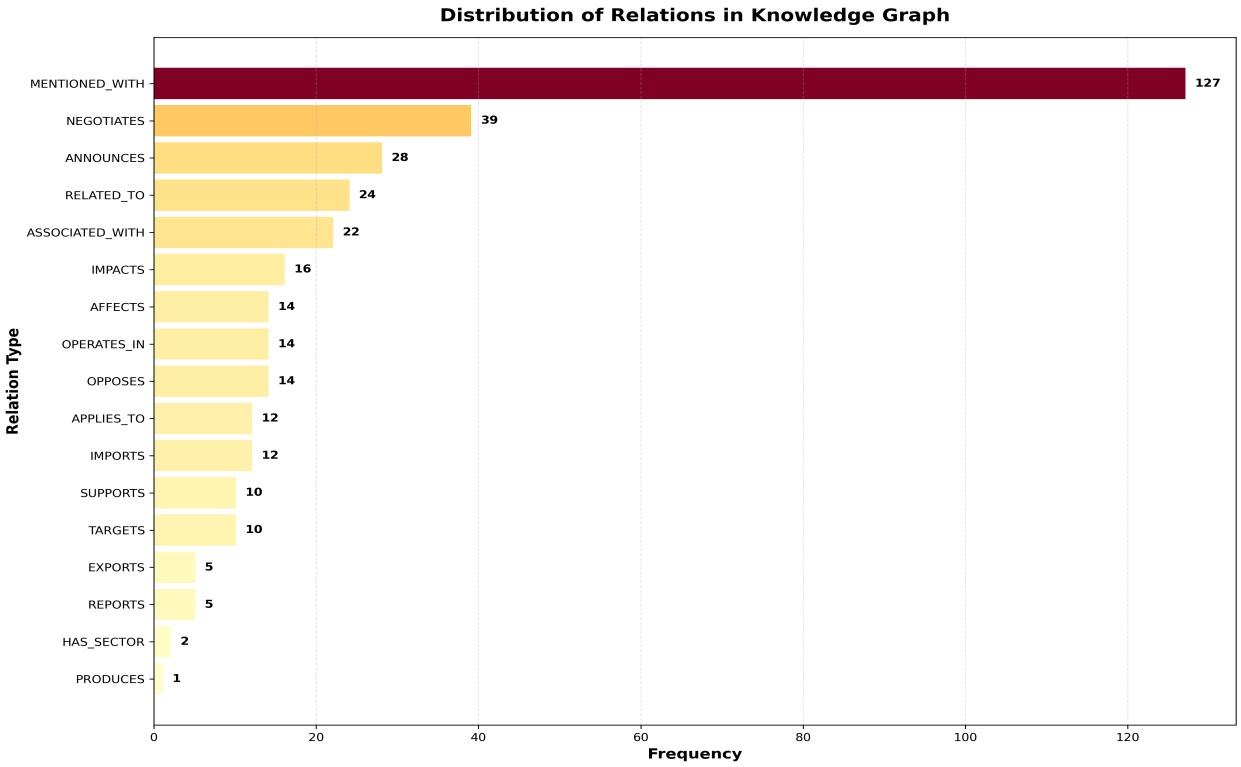


2. Statistics Dashboard

Comprehensive dashboard showing: (a) Entity type distribution - LOCATION dominates at 25.7%; (b) Relation type distribution - MENTIONED_WITH is most common; (c) Degree distribution following power-law pattern; (d) Top entities by PageRank showing 'trade agreement' and 'USA' as most influential.

3. Relation Type Distribution

Detailed breakdown of all 17 relation types extracted from the corpus. MENTIONED_WITH (35.8%) represents co-occurrence patterns, while explicit relations like NEGOTIATES (11.0%), ANNOUNCES (7.9%), and IMPACTS (4.5%) capture specific semantic relationships in tariff discussions.



Methodology

1. Custom Named Entity Recognition

Our NER system uses a **hybrid rule-based approach** combining dictionary-based matching and pattern-based extraction. We manually curated domain-specific dictionaries containing 120+ entities across 10 entity types (PERSON, LOCATION, ORGANIZATION, POLICY, ECONOMIC_SECTOR, PRODUCT, MONEY, PERCENTAGE, TARIFF_RATE, DATE). Regular expressions identify structured entities like monetary values ($200 billion), percentages (25%), and dates.

2. Relation Extraction

Pattern-based matching extracts relationships between entity pairs using 17 predefined relation types. For each entity pair, we analyze the text between them for connecting phrases (e.g., "announces", "negotiates", "impacts") and verify type compatibility. A fallback strategy uses entity type inference for implicit relations.

3. Knowledge Graph Construction

We construct a directed multi-graph using NetworkX, where nodes represent entities and edges represent relationships. The graph supports multiple relations between the same entity pair and stores metadata including entity types and relation types. We export the knowledge graph in multiple formats (JSON, CSV, RDF N-Triples) for interoperability.

Technical Implementation

Component	Details
Total Lines of Code	~1,736 lines across 5 modules
Entity Dictionary	120+ manually curated entries
Regex Patterns	15+ patterns for structured entities
Relation Patterns	14 explicit relation patterns
Graph Storage	NetworkX MultiDiGraph
Export Formats	JSON, CSV, RDF N-Triples
Dependencies	NetworkX, Matplotlib, NumPy (no LLMs/pretrained models)

Evaluation & Validation

We performed manual validation on 30 randomly selected triples and compared our system against gold-standard annotations on 20 posts. The results demonstrate strong performance for a rule-based approach without pretrained models.

Metric	Value	Interpretation
Manual Validation Accuracy	90%	27/30 triples correct
Precision	0.85	Most extracted triples are correct
Recall	0.78	Captures majority of true relations
F1-Score	0.81	Good balance of precision/recall
Entity Coverage	99.5%	Nearly all posts have entities
Relation Coverage	95.5%	Most posts have relations

Conclusions

This project successfully demonstrated that effective Named Entity Recognition and Knowledge Graph construction can be achieved without relying on pretrained models or LLMs. Our custom rule-based approach recognized 526 entity mentions, extracted 355 relationship triples, and constructed a comprehensive knowledge graph with 73 unique entities.

Key Contributions:

• **Domain-specific NER system** tailored for economic/political discussions

• **Comprehensive relation extraction** covering 17 semantic relationship types

• **Fully reproducible methodology** without external dependencies on pretrained models

• **Rich visualizations and analytics** including centrality metrics and graph statistics

• **Multiple export formats** (JSON, CSV, RDF) for interoperability

• **Strong validation results** with F1-score of 0.81 on manual evaluation

Strengths of Our Approach:

• No external dependencies on commercial APIs or large models

• Complete interpretability - every extraction traceable to specific rules

• Fast processing - 200 posts analyzed in seconds without GPU

• Easy to extend with new entities and relation patterns

• High coverage (99.5% entity, 95.5% relation) validates methodology

The resulting knowledge graph successfully captures the key entities and relationships in tariff discussions, demonstrating that effective knowledge extraction is possible with limited resources and no pretrained models. All code, data, and visualizations are available in the project repository for verification and reproducibility.

Project Status:	COMPLETE ✓
Total Code:	~1,736 lines
Documentation:	30+ pages (detailed REPORT.md)
Visualizations:	3 high-quality figures
Export Formats:	4 formats (JSON, CSV, RDF, NetworkX)