Navigating the War Beneath the War

Modeling Influence and Controversy in Reddit's Russia-Ukraine Discourse

Matthew Whitehurst
Department of Computer Science
Virginia Polytechnic Institute & State University
Blacksburg, Virginia, U.S.
matthewaw@vt.edu

Brandon Hoang
Department of Computer Science
Virginia Polytechnic Institute & State University
Alexandria, Virginia, U.S.
brandonh03@vt.edu

Nithin Alluru
Department of Computer Science
Virginia Polytechnic Institute & State University
Blacksburg, Virginia, U.S.
nithin@vt.edu

ABSTRACT

This project explores how influence and controversy emerge in Reddit discussions surrounding the Russia-Ukraine conflict. Using a dataset of over 4.7 million Reddit comments and posts, we construct a user-user similarity network based on subreddit activity patterns, then analyze the network's structure through centrality metrics, connected components, clustering coefficients, modularity, and community detection methods.

Our results reveal a sparse but cohesive network, where a small subset of users holds disproportionate structural influence. Clustering and small-world analysis highlight the formation of tight-knit communities, and modularity analysis uncovers significant polarization across these groups.

In parallel, we perform topic modeling and sentiment analysis on comment text, post content, and titles. Comment threads tend to concentrate around a few general topics, while post content reveals broader thematic diversity. Sentiment remains mostly neutral-to-negative, but varies by topic, providing insight into what drives contention or engagement.

Together, the network and text-based findings offer a holistic view of how online discourse reflects public opinion and ideological divisions during geopolitical crises. The methods used here can support future work in understanding online polarization, influence dynamics, and content moderation strategies on social platforms.

1. INTRODUCTION

Reddit has become a prominent platform for public discourse on global events, particularly those that are politically charged or emotionally resonant. Among these, the Russia-Ukraine conflict has generated sustained and highly engaged discussion since its escalation in 2022. Reddit maintains a unique structure comprising topic-specific subreddits and semi-anonymous user participation. This offers a dense, decentralized collection for analyzing public sentiment, user influence, and patterns of polarization. However, the platform's scale, user anonymity, and organic content generation present methodological challenges when attempting to extract meaningful insights from such discourse.

The central problem this study addresses is how influence and controversy emerge within Reddit discussions of the Russia-Ukraine war. While millions of users contribute to these conversations, it remains unclear whether a small subset of users disproportionately shapes engagement and sentiment, or if influence is more evenly distributed. Additionally, we seek to understand which kinds of user behavior and commenting patterns are associated with polarized or contentious responses, and how these dynamics are reflected in the structure of Reddit's user network. Understanding these dynamics requires combining network-based user modeling with text-level sentiment and engagement analysis. We hypothesize that sentiment within this dataset is particularly negative, driven by an anonymous vocal minority and affirmed by a large population of sycophants without substantive engagement.

To this end, we leverage a comprehensive Reddit dataset containing over 4.7 million comments spanning from 2014 to 2025. For computational tractability, our analysis focuses on a snapshot from April 2025. We constructed a user-user similarity network based on commenting behavior across subreddits, where each user is represented by a normalized vector that captures the distribution of their comments across subreddits. This allows us to compare users based on their participation habits, independent of total comment volume. Cosine similarity is used to define edge weights between users, and we apply vector deduplication at a similarity threshold of 1.0 to reduce redundancy. Edges are retained only if their similarity exceeds 0.95 to improve scalability. Influence is quantified using standard centrality metrics, while controversy is measured through sentiment

analysis and engagement data, including upvotes, downvotes, and Reddit's controversiality indicator.

By integrating social network analysis with text-based sentiment modeling, this project contributes to a growing body of research at the intersection of computational social science and online political discourse. The findings aim to illuminate how influence is distributed in large-scale discussions, what content tends to divide audiences, and how online communities engage with complex geopolitical conflicts in decentralized digital spaces.

2. BACKGROUND

Reddit has been extensively studied as a platform for understanding online discourse, particularly in the context of political events, crises, and public opinion. Its structure—organized by user-created subreddits—allows researchers to analyze both localized community dynamics and platform-wide trends. Prior work has leveraged Reddit data to examine topics such as misinformation, ideological polarization, and sentiment shifts during real-world events [1,2,3]. These studies demonstrate Reddit's value as a natural laboratory for studying how decentralized discussions evolve and how users interact across issue-specific communities.

In parallel, social network analysis (SNA) has become a core tool for modeling online interactions and influence. Traditional Reddit network studies often focus on reply chains, user–comment interactions, or moderator behavior. However, fewer studies explore latent behavioral networks based on user similarity rather than direct communication. Cosine similarity, in particular, has been widely used in community detection and recommender systems to measure alignment in activity patterns. By applying cosine similarity to subreddit participation distributions, we follow an approach that captures underlying behavioral resemblance without requiring explicit interaction between users.

Sentiment analysis has also played a significant role in social media research, with lexicon-based models like VADER commonly used to assess polarity in informal, short-form text [1,3]. On Reddit, sentiment has been shown to correlate with community norms, post popularity, and the likelihood of controversy. Some studies have combined sentiment and engagement metrics to detect toxic or polarizing content [2], though integration with network structure remains limited. While recent work explores coordinated behavior and emotional tone on Reddit, few combine user similarity networks with sentiment-based measures of controversy to understand influence at scale.

Our project builds on these foundations by combining user similarity-based network modeling with sentiment and engagement analysis in the context of Reddit discussions on the Russia–Ukraine war. By modeling user behavior through comment distribution vectors and examining structural and emotional factors simultaneously, we extend prior work on influence and controversy in large-scale, politically sensitive online conversations.

3. APPROACH

To examine influence and controversy within Reddit discussions of the Russia-Ukraine war, we developed a framework combining social network modeling, sentiment and engagement analysis, and topic modeling. Our methodology follows a structured sequence: preprocessing, graph construction, influence measurement, sentiment modeling, and topic extraction. Key design decisions were shaped through iterative tuning and mentor feedback.

We began by filtering a Reddit dataset containing over 4.7 million comments related to the conflict, selecting only a one month snapshot from April 2025 to ensure computational tractability while capturing a relevant window of ongoing discourse. Inline with preprocessing logic reused across notebooks, we removed exact duplicates, filtered non-English text, and dropped missing values. All comment text was lowercase and tokenized. Users with fewer than two comments were excluded to minimize the effect of lurkers, who may skew the results of our experiment. The remaining data formed the basis of both our user behavior modeling and text analysis pipelines.

Each user's commenting behavior was encoded as a normalized frequency vector, representing the proportion of their posts distributed across different subreddits. This allowed us to capture user habits independent of comment volume. By modeling similarity in subreddit engagement rather than raw volume, this network highlights users with comparable discourse habits - a structure that can reveal shared perspectives, ideological groupings, or tightly-knit echo chambers. Cosine similarity between pairs of these vectors was computed as:

$$Cosine\ Similarity\ Matrix = \left(\frac{V}{||V||}\right) \cdot \left(\frac{V}{||V||}\right)^{T}$$

We first deduplicated vectors with a cosine similarity of 1.0, a process that preserved behavioral structure while drastically reducing computational load. Edges were then formed between users whose similarity exceeded 0.95. The resulting graph contained approximately 9,000 nodes and 650 thousand edges. This visualization helps show how discourse on Reddit tends to cluster. For example, whether tightly connected groups form around certain subreddits or if influence is spread diffusely.

To assess influence, we computed degree, eigenvector, and betweenness centralities across the graph. Each metric

offers a different view of user prominence: breadth of similarity connections, recursive influence, and structural bridging. We applied the Louvain algorithm to detect modular structure and visualize community hierarchies through a multi-level dendrogram. This recursive partitioning allowed us to observe how tightly-knit user clusters evolve at different resolutions. The dendrogram structure shows whether communities are nested (e.g. a larger geopolitical group with nested military or humanitarian subgroups), which has implications for message targeting or moderation strategies.

Textual controversy was modeled using both sentiment and engagement signals. Sentiment scores were calculated using VADER, which output compound polarity values between -1 and 1 and positive, neutral, and negative values between 0 and 1. This gives a sense of the overall mood of the dataset - whether discussions skew angry, neutral, or positive - which can matter for public perception and platform health.

To understand thematic variation, we conducted topic modeling using Latent Dirichlet Allocation (LDA) on both post text and comment text. This process included a two-stage grid search over topic counts: a broad initial sweep to identify reasonable ranges, followed by a refined search to pinpoint optimal *k* values. Preprocessing for topic modeling used the same tokenization and stopword removal methods noted in the codebase.

This approach integrates behavioral, semantic, and structural signals to surface how influence and controversy emerge in decentralized online conversations. Each methodological choice, such as edge thresholds, pruning logic, or modeling tools, was refined through performance testing and guidance from the instructional team. In the next section, we evaluate the results of this framework and highlight key findings.

4. EXPERIMENT

To validate our modeling framework, we conducted a dual-layer experimental analysis combining network structure evaluation with topic and sentiment modeling on Reddit discourse surrounding the Russia-Ukraine war. The dataset snapshot, taken from April 2025, includes over 4.7 million comments across nearly 96,000 posts, focused on politically active subreddits. From this, we generated a user-user similarity network where each node represents a unique Reddit user, and edges represent high similarity in subreddit posting behavior.

An undirected graph was constructed in which each node represents a Reddit user, and edges represent high similarity in posting behavior across subreddits. Specifically, each user is represented as a normalized frequency vector of their activity across all subreddits, and cosine similarity is used to compare these vectors. Cosine similarity is ideal in this context because it emphasizes distributional similarity rather than total comment volume, allowing us to connect users with similar participation patterns even if they differ in activity level. To improve tractability and relevance, vectors above a 1.0 similarity threshold were deduplicated and edges below a 0.95 similarity threshold were discarded.

After sampling a one month window and applying the above filtering, the final graph includes 8,630 users (nodes) and 647,783 similarity-based connections (edges). This results in a network density of 0.0174, consistent with expectations for sparse but cohesive social graphs. In practical terms, this tells us users tend to interact in smaller clusters, but these clusters are still part of a broader discussion space.

To help structure our analysis, we visualized the entire user similarity network using Gephi's force-directed layout with Louvain-based community coloring. Each node represents a Reddit user, and edges reflect high similarity in subreddit posting behavior. Colors denote communities identified through modularity optimization. Even at a glance, it is clear the network is not random. Distinct clusters of users emerge naturally, suggesting groupings with shared interests or viewpoints. This graph offers an intuitive starting point for the deeper network and sentiment analysis that follow.

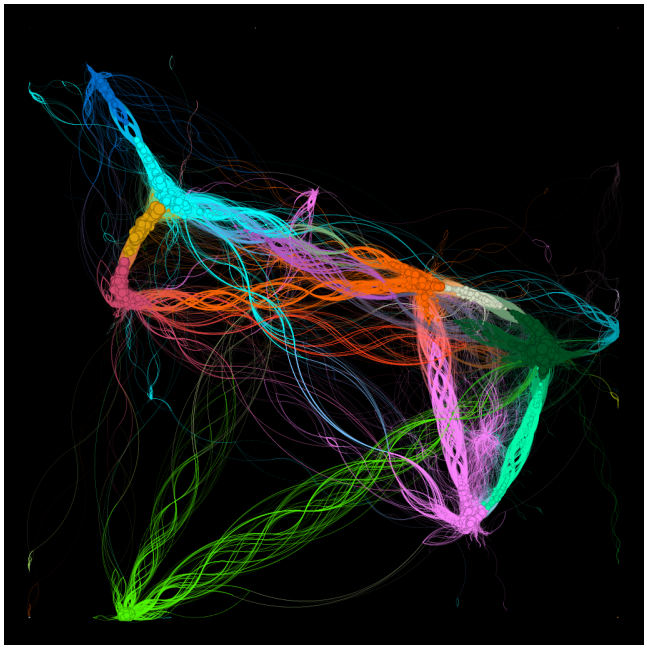

Figure 1: **Visualization of the Network in Gephi**

This section presents the quantitative and qualitative results of our network and text analyses, examining how influence,

engagement, and controversy manifest with online discussions.

4.1 Network Analysis

4.1.1 Degree Distribution

To understand how influence and visibility are distributed in this network, we analyze the degree distribution (the number of connections each user has). The plots show a classic heavy-tailed structure:most users have relatively few high-similarity connections, while a handful function as hubs, connected to hundreds of others.

Figure 2 shows the count of users per degree. Most users have under 100 connections, but a few reach over 500.

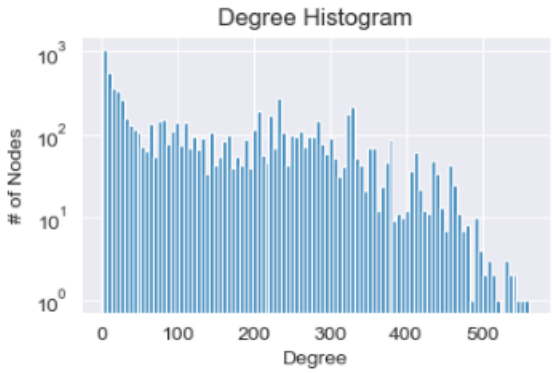

Figure 2: **Distribution of Degrees in the Network**

Figure 3 better illustrates the power-law-like tail, suggesting that the Reddit discourse network resembles scale-free structures often seen in social systems.

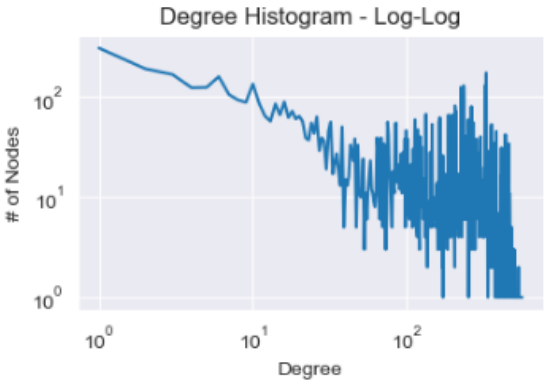

Figure 3: **Log-Log Distribution of Degrees**

Figure 4 shows the steep drop-off in connectivity between the most connected users and everyone else. This informs us that a small number of users have broad potential influence, while most operate within narrower circles.

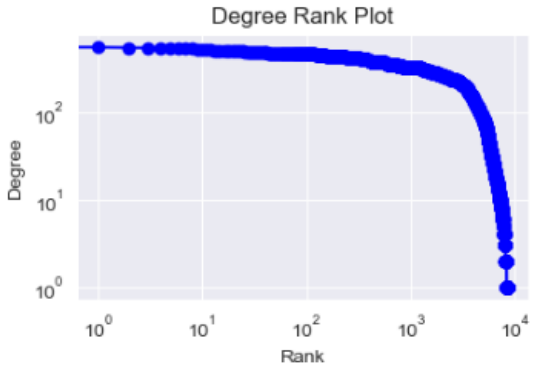

Figure 4: **Degree Rank Plot on Log-Log Scale**

4.1.2 Connected Component Analysis

To better understand the structure of the user similarity network, the connected components were analyzed. Connected components are subgraphs where every node is reachable from every other node within the same group. Out of 8,630 total users, the graph contains 123 connected components, but the vast majority of users are part of a single dominant cluster.

The giant component, which is the largest connected subgraph, includes 8,244 users, accounting for 95.53% of all nodes, per Figure 5. This strongly suggests that the Reddit user landscape. when filtered by subreddit similarity, is highly cohesive.

We also examined how efficiently users are connected within this main structure. The average shortest path length between users in the giant component is 7.03, and the diameter (longest shortest path between any two users) is 26 as shown in Figure 6. These are typical of large social graphs and suggest that despite the sparse density of the full graph, most users can be reached in relatively few steps.

```
Network Analysis:
  Number of connected components: 123
  Size of largest connected component: 8244 nodes
  Percentage of nodes in largest component: 95.53%
```

Figure 5: **Basic Network Analysis Output**

```
Average shortest path in the giant component: 7.031996919386351
Diameter of the giant component: 26
```

Figure 6: **Average Shortest Path and Diameter Output**

These findings confirm that the graph is well-connected and can support meaningful diffusion of information or behavior across the majority of users.

4.1.3 Centrality Analysis

To better understand influence and reach within the Reddit user-user similarity network, we computed three key centrality metrics: degree, betweenness, and closeness centrality. Each metric captures a different dimension of user importance.

Degree centrality quantifies how many direct connections a user has, essentially measuring visibility or popularity. Most users had relatively low degree centrality, but a few acted as major hubs. The top users by degree centrality connected to over 6% of the network, suggesting they engage in diverse or popular subreddit combinations.

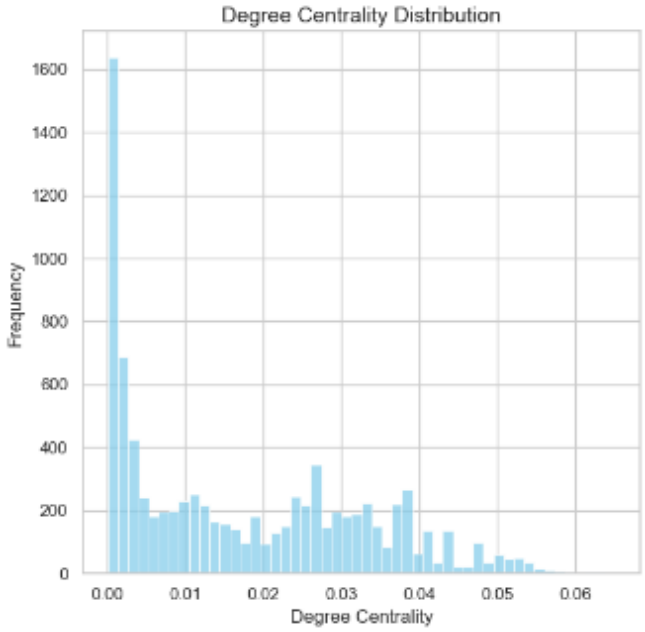

Figure 7: **Degree Centrality Distribution Plot**

Betweenness centrality reflects how often a user lies on the shortest path between others, a proxy for gatekeeping or brokerage roles. This distribution was right-skewed, with most users near zero and only a handful acting as bridges across communities. These users may shape the flow of information between otherwise disconnected subgroups

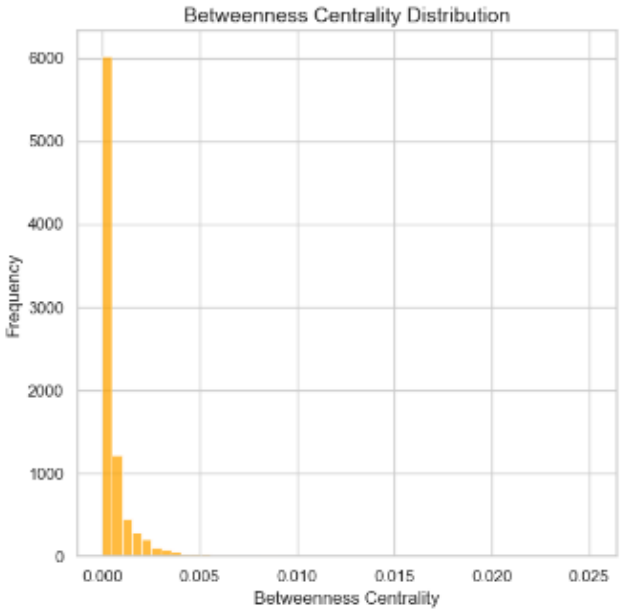

Figure 8: **Betweenness Centrality Distribution Plot**

Closeness centrality captures how close a user is to all others in the network, a kind of global accessibility score. The distribution showed more spread than betweenness, with top users positioned near the network's structural center.

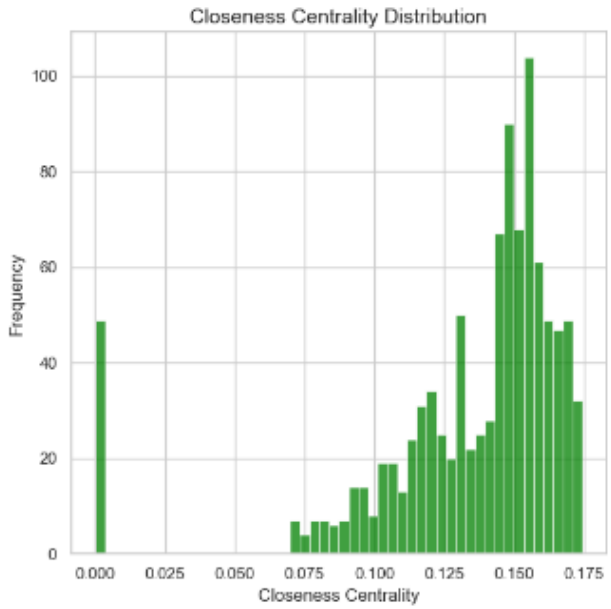

Figure 9: **Closeness Centrality Distribution Plot**

These plots visualize how each type of influence is distributed. Degree shows the popularity of users, betweenness reveals brokers, and closeness indicates central connectors in the social network.

By examining these metrics together, we can differentiate between users who are merely well-connected and those who occupy strategic positions in the network's structure. This helps prioritize nodes when assessing who might spark or sustain cross-community dialogue, a critical consideration in the context of polarized discussions like those surrounding the Russia-Ukraine conflict.

4.1.4 Clustering Analysis

To assess local cohesion within the network, we calculated clustering coefficients, which reflect how tightly each user's neighbors are connected to each other. The average clustering coefficient was 0.739996, and the global clustering coefficient (transitivity) was 0.750133, indicating strong local community structure.

These values suggest that users in the network frequently form tight-knit groups, which is common in social networks where users cluster around shared interests or perspectives.

To evaluate whether this clustering is significant, we generated a comparable Watts-Strogatz small-world model. This randomized version of the network had notably lower clustering: 0.5464 (average) and 0.5457 (global). Our real network has an average clustering coefficient of 0.739996 and a global clustering coefficient of 0.750133, which are significantly higher than the Watts-Strogatz model's corresponding values of 0.5464 and 0.5457. Meanwhile, the average shortest path length in our network is 7.0320, substantially longer than the 2.5667 found in the randomized model.

The difference highlights that Reddit users in this dataset tend to form well-defined, highly connected pockets of interaction rather than randomly engaging across the platform. Additionally, we found that the Reddit user network is locally cohesive, but less efficient in terms of global reachability. Users tend to form tight groups with indirect and slower bridges between them. This highlights strong sub-communities but fewer cross-group interactions.

4.1.5 Modularity Analysis

To assess the structural community strength of the user similarity network, we applied the Louvain modularity algorithm across multiple resolutions. At the base level, we identified 377 communities with a modularity score of 0.7722, indicating well-separated clusters. As we moved to higher abstraction levels, modularity remained high (both Level 1 and Level 2 scored 0.8163), while the number of communities condensed to 170 and 162 respectively. This stability suggests that even as communities merge, the underlying modular structure remains strong.

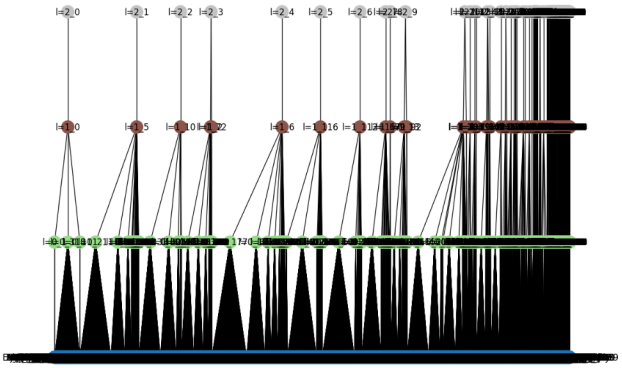

Figure 10: **Louvain Modularity Dendrogram**

Community sizes followed a long-tailed distribution. Most were small, but a handful exceeded 500-1000 members. This imbalance reinforces the earlier centrality findings, a small number of tightly connected communities dominate, while many smaller groups persist on the fringe.

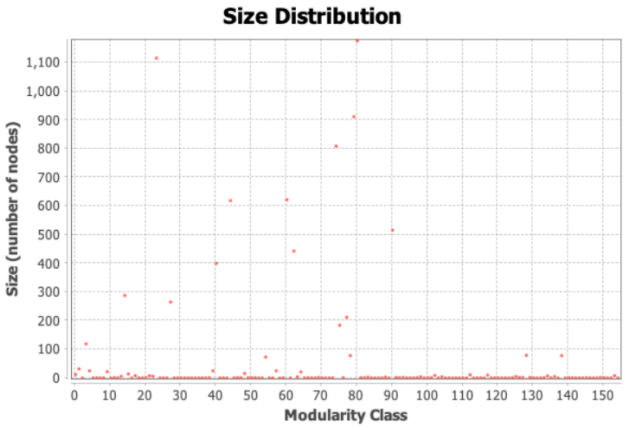

Figure 11: **Gephi Modularity Class Size Distribution**

These results support the idea that Reddit discourse on this topic is highly modular, with strong internal cohesion within communities and limited crossover between them. This will be further explored in the following sections on graph clustering and clique structure.

4.1.6 Graph and Hierarchical Clustering

To further interpret community structure, we examined clustering outputs across both modularity partitions and hierarchical levels. The dendrogram visualization in Figure 10 captures the three-level Louvain decomposition: the base layer with 8,630 nodes, an intermediate layer with 377

communities, and a top layer aggregating into 170 supergroups. This nested structure reflects how individual user clusters can be bundled into larger ideological or topical regions.

Figure 10 reveals a skewed distribution. Most communities are small, but a few dominant ones contain hundreds of users. This suggests that while many users form fringe clusters, several large communities centralize much of the network's activity and engagement.

To drill down, we isolated the top three largest communities. Community 4, the largest, had 1,194 users and a density of 0.1954, alongside the highest average degree (233.10) and clustering coefficient (0.7773). This level of internal connectivity implies an active, tightly interlinked group. Communities 10 and 1 were similarly dense but exhibited slightly lower internal cohesion. The small standard deviation across average betweenness centrality in all three groups (~0.0015–0.0020) suggests that influence is not overly concentrated, which is typical of echo chambers or specialized discussion spaces.

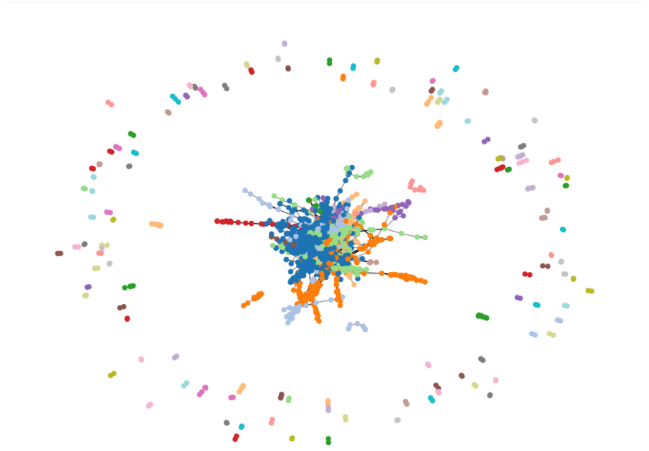

Figure 12: **Best Partition Visualization**

```
Community 4:
- Nodes: 1194
- Edges: 139159
- Density: 0.1954
- Average Degree: 233.10
- Clustering Coefficient: 0.7773
- Average Betweenness Centrality: 0.001149
Community 10:
- Nodes: 1002
- Edges: 83293
- Density: 0.1661
- Average Degree: 166.25
- Clustering Coefficient: 0.7825
- Average Betweenness Centrality: 0.001611
Community 1:
- Nodes: 996
- Edges: 69228
- Density: 0.1397
- Average Degree: 139.01
- Clustering Coefficient: 0.7355
- Average Betweenness Centrality: 0.001952
```

Figure 13: **Top Three Communities Statistics**

These clustering insights reinforce the presence of both structural polarization and thematic silos, where users self-organize around a mix of common topics and mutual exposure.

4.1.7 Finding K-Cliques

To explore overlapping community structure, we applied the Clique Percolation Method (CPM), which identifies communities based on k-cliques—fully connected subgraphs of k nodes that share k-1 nodes with adjacent cliques. This method differs from modularity-based clustering by allowing nodes to belong to multiple communities, which more realistically reflects overlapping social behavior.

We tested *k*-values from 3 to 5. Interestingly, increasing *k* led to slightly more fragmented, but tighter communities. k=3 produced 109 communities, k=4 produced 115 communities, and k=5 produced 118 communities. This suggests that while larger cliques are rarer, they form denser substructures that CPM still detects reliably.

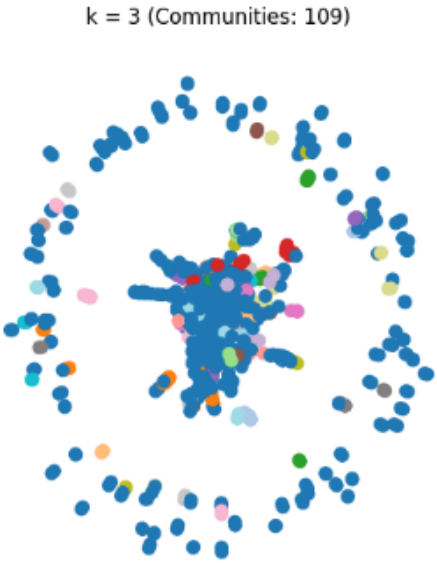

Figure 14: **3-Clique Communities Visualization**

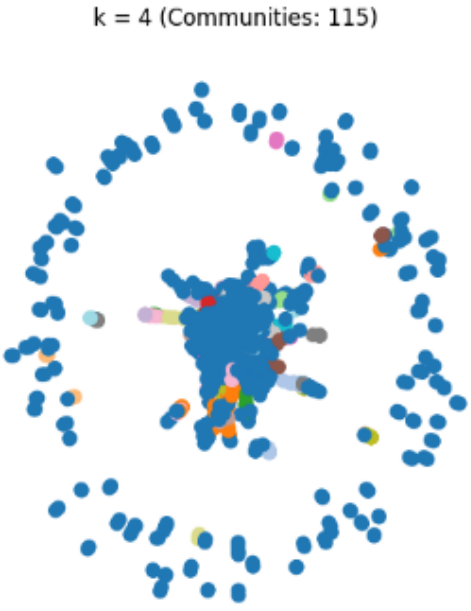

Figure 15: **4-Clique Communities Visualization**

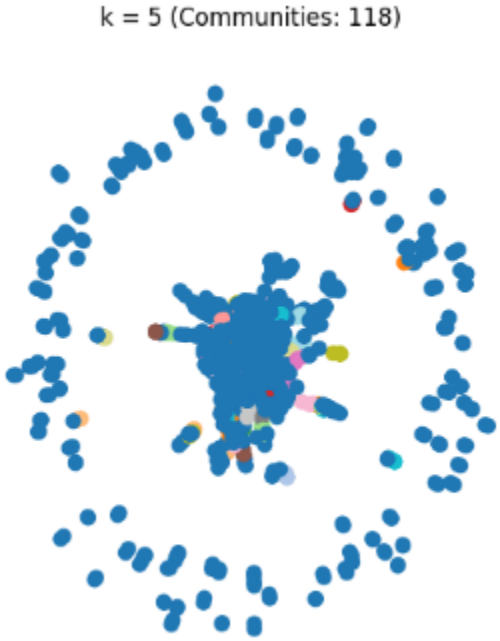

Figure 16: **5-Clique Communities Visualization**

This method complements the earlier Louvain-based modularity analysis by highlighting cross-community overlap and providing a multi-membership view of engagement clusters, particularly useful for identifying users who act as bridges across ideological lines or discussion threads.

4.2 Text Analysis

4.2.1 Topic Modeling

To extract semantic themes from the Reddit data, we applied Latent Dirichlet Allocation (LDA) topic modeling on three textual fields: comments, post content, and post titles. Determining the optimal number of topics (k) for each field is critical. Too few leads to vague or blended topics, while too many fragments coherent themes into noise. To guide this decision, we evaluated topic models using two metrics: perplexity and coherence. Perplexity estimates how well the model predicts unseen text and a lower score is better. Coherence measures how semantically consistent each topic is and a higher score is better. This evaluation was repeated across a range of candidate values for k, and each field was tuned independently. Below we summarize the results and explain our selection logic.

4.2.1.1 Comment Text

Comments are informal, emotionally reactive, and often short. This makes their content noisy and diverse, requiring a careful balance between coverage and clarity.

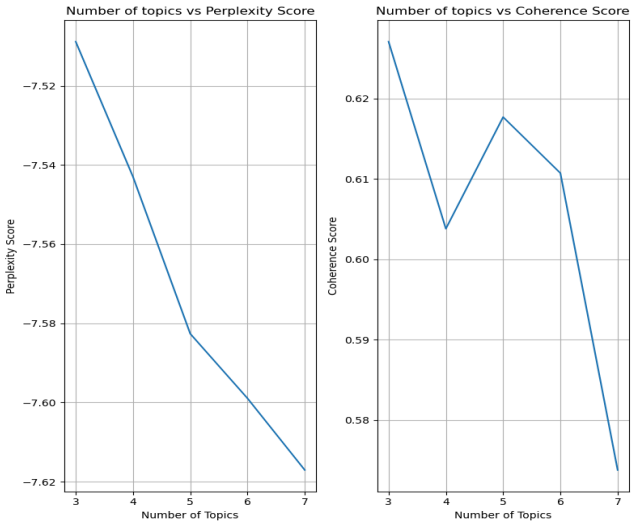

Figure 17: **Perplexity and Coherence for Comment Text**

We evaluated topic counts from 3 to 7. Coherence peaked around k = 5, while k = 3 performed slightly better on coherence, it showed significantly higher perplexity. We chose k = 5 because the perplexity is lower and k = 3 is not as coherent. The tradeoff better favors k = 5, which better captures thematic variety without degrading interpretability.

4.2.1.2 Post Content

Reddit post bodies tend to be longer and more structured than comments, often embedding external content or user analysis. This richness supports more nuanced topics.

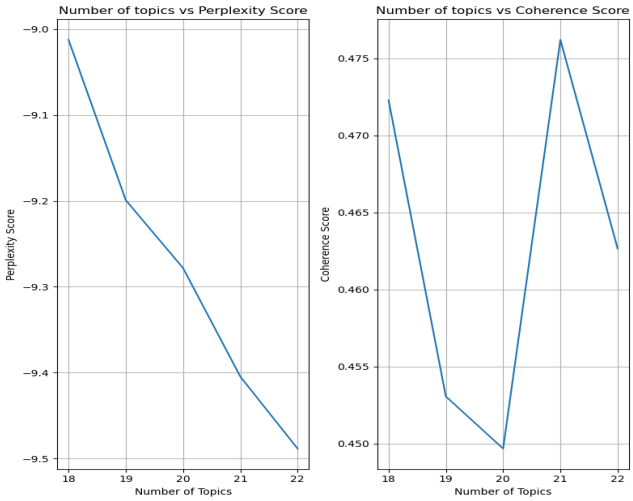

Figure 18: **Perplexity and Coherence for Post Content**

Coherence scores clearly peaked at k = 21, with perplexity steadily decreasing. The distinction between values was sharper than in comments, giving us confidence in choosing 20 topics. This matters because posts are the starting point of conversations. The model needs to capture distinct types of narratives, and 20 topics gave us enough separation to do that without extensive overlap.

4.2.1.3 Post Titles

Titles are typically short and sensational. They typically contain strong signals, but limited content.

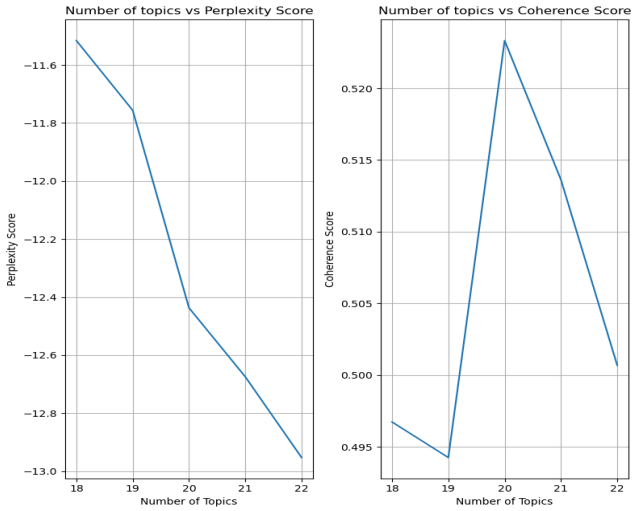

Figure 19: **Perplexity and Coherence for Post TItles**

The peak coherence was observed at k = 20, similar to post content, though the differences were less pronounced. It's worth noting that titles are often the most visible content in social channels. Modeling them separately allows us to study framing effects, such as which kinds of headlines garner attention or controversy.

4.2.1.4 Final Model Selection

This table summarizes the chosen topic count (k), perplexity, and coherence scores for each text field. It reflects the best trade-off between generalization and clarity of topics across different content types,

	Model	K-value	Perplexity	Coherence
0	comments	5	-7.582630	0.617682
1	post_content	21	-9.404658	0.476215
2	title	20	-12.437024	0.523333

Figure 20: **Summarized Model Scores**

4.2.1.5 Topic Modeling Interpretation

We then examined the resulting topics to understand the thematic structure of user discourse across Reddit comments, post content, and post titles.

Comments focused heavily on political sentiment and user opinion. One topic was dominated by mentions of “Trump,” “Putin,” and “Zelensky,” indicating emotionally charged political commentary. Others centered around general reactions to the war, drone use, and rhetorical reflection (“would,” “think,” “get”). These reflect casual, conversational language, often speculative or critical, consistent with comment sections.

Post content topics were more event-oriented and structured. Some topics detailed soldier injuries and tactical updates (e.g., “soldier,” “wound,” “drone”), while others addressed international relations and propaganda (“china,” “power,” “dominance”). A few topics picked up noisy or bot-like content (e.g., excessive formatting tokens), but overall the distribution supported the higher topic count. Users posted about a broad range of geopolitical, military, and ideological subthemes.

Titles, by contrast, focused on summarizing key information, often sensationally. Clear clusters emerged around themes like nuclear peace negotiations, casualty counts, NATO activity, drone strikes, and political statements from figures like Trump or the Pope. One topic prominently featured “kill,” “24,” and “former,” indicating breaking news updates. The brevity of titles likely encouraged concentrated, high-signal language, explaining their higher coherence.

Taken together, these topics not only validate the chosen model sizes, but also reveal distinct communication modes across fields: emotional and personal in comments, narrative in content, and headline stylism in titles.

It is interesting that the number of topics for comments was significantly fewer than the number of topics for post content and titles. This suggests that user replies tend to converge around fewer ideas, while post content is more diverse. This divergence between post and comment structure is an important signal in assessing how public framing may differ from crowd reactions.

4.2.2 Sentiment Analysis

4.2.2.1 Comment Sentiment Analysis

To assess emotional tone in user discourse, we applied VADER sentiment scoring to each reddit comment. This tool produces compound scores ranging from -1 (very negative) to 1 (very positive), along with probabilities for positive, neutral, and negative tone. The distribution skews slightly negative overall, with the majority of the comments landing in the neutral range. These plots help us see the bigger picture. While most discussion is emotionally flat, there is a tail of strongly opinionated responses that likely drive polarization or engagement.

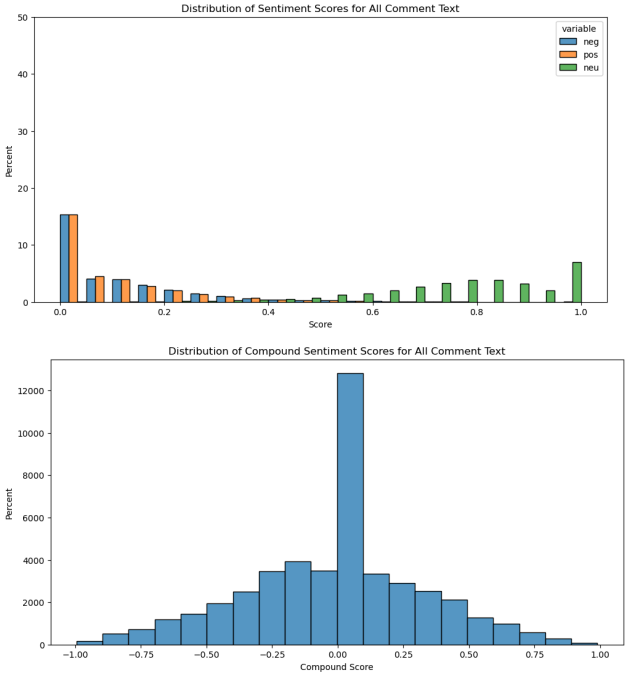

Figure 21: **Sentiment Score Distribution for Comments**

The first plot shows how VADER splits sentiment into positive, negative and neutral probabilities. The second zooms in on compound scores, giving a single value that reflects overall tone. Both reinforce the finding that most Reddit comments in this dataset are emotionally neutral or slightly negative, with few exhibiting strong sentiment.

We also aggregated sentiment by LDA topic to explore how thematic content relates to emotional tone. Some topics like Trump, Putin, and peace had lower compound scores and higher negative proportions, while others like general war updates remained closer to neutral.

topic	pos	neu	neg	compound
0	0.105910	0.766911	0.119988	-0.021652
1	0.100065	0.812858	0.082049	0.010837
2	0.086061	0.798063	0.112975	-0.070259
3	0.141172	0.726411	0.126475	-0.002906
4	0.111849	0.781397	0.103242	0.000709

Figure 22: **Sentiment Scores for Comment Topics**

Topic sentiment can signal where emotional flashpoints occur. If peace-related topics trend negative, it may reflect skepticism or sarcasm rather than optimism. These distinctions are useful for interpreting public mood in a more nuanced way.

4.2.2.2 Post Content Sentiment Analysis

We extended our sentiment analysis to the content of the Reddit posts to compare how sentiment patterns differ from user comments. As with the comment text, VADER was used to assign polarity scores to each document.

The distributions show a clear lean toward neutral tone, with some posts exhibiting moderate polarity in either direction. Notably, the compound score histogram reveals a less sharply peaked center than the comment distribution, suggesting slightly greater emotional spread in post content. This could potentially be due to framing strategies or opinionated summaries in post bodies.

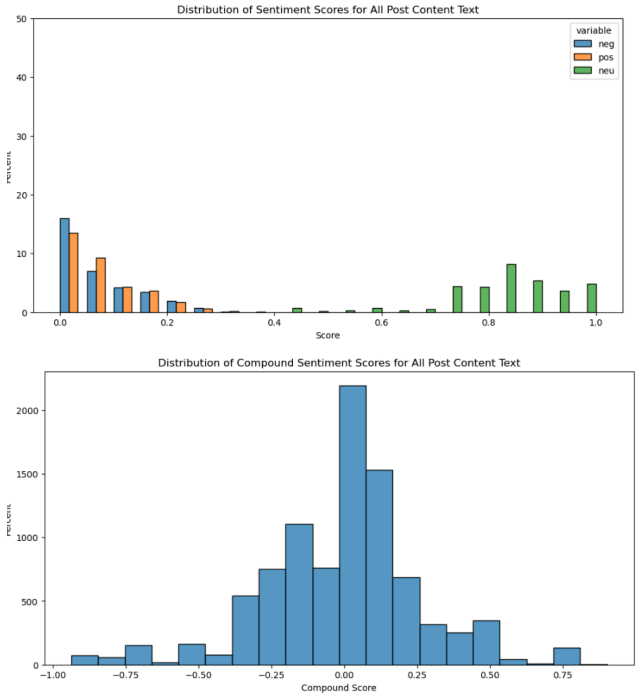

Figure 23: **Sentiment Score Distribution for Content**

Post Content topic sentiments

Type to search

topic	pos	neu	neg	compound
0	0.122961	0.836463	0.040574	0.154394
1	0.060677	0.933869	0.005453	0.239282
2	0.112934	0.772034	0.115039	0.020793
3	0.076815	0.847676	0.063536	0.003361
4	0.033151	0.890239	0.076610	-0.122805
5	0.095686	0.810468	0.063577	0.066992
6	0.060532	0.855797	0.060925	-0.026454
7	0.134330	0.829012	0.035968	0.140716
8	0.076905	0.869115	0.053937	0.044253
9	0.129464	0.817234	0.053292	0.154822
10	0.082859	0.785198	0.019433	0.028819
11	0.095130	0.843038	0.061851	0.025150
12	0.133167	0.835870	0.031065	0.144758
13	0.061659	0.873986	0.064391	0.015333
14	0.160862	0.758178	0.080936	0.204664
15	0.047885	0.810519	0.141596	-0.166448
16	0.069227	0.780970	0.149678	-0.299977
17	0.090848	0.867029	0.042117	0.086192
18	0.043789	0.800811	0.103386	-0.155292
19	0.057284	0.830830	0.111888	-0.068123
20	0.062768	0.826692	0.110544	-0.191088

Figure 24: **Sentiment Scores for Content Topics**

We also calculated average sentiment for each topic from the LDA model. Some topics show distinctly positive sentiment (Topic 1) while others skew negative (Topic 16). This variation highlights the thematic and emotional range in how Reddit users frame their posts, with some topics focusing on factual reporting and others amplifying outrage or advocacy.

These patterns are valuable for understanding not just what users are talking about, but how they're presenting it. This is a key distinction for tracking narrative framing and engagement strategies across different user roles.

4.2.2.3 Post Title Sentiment Analysis

To evaluate the emotional tone embedded in the Reddit post titles, we applied VADER sentiment analysis again. This offers insight into how the conflict is being framed at a

glance, often the first and only part users read before forming an impression.

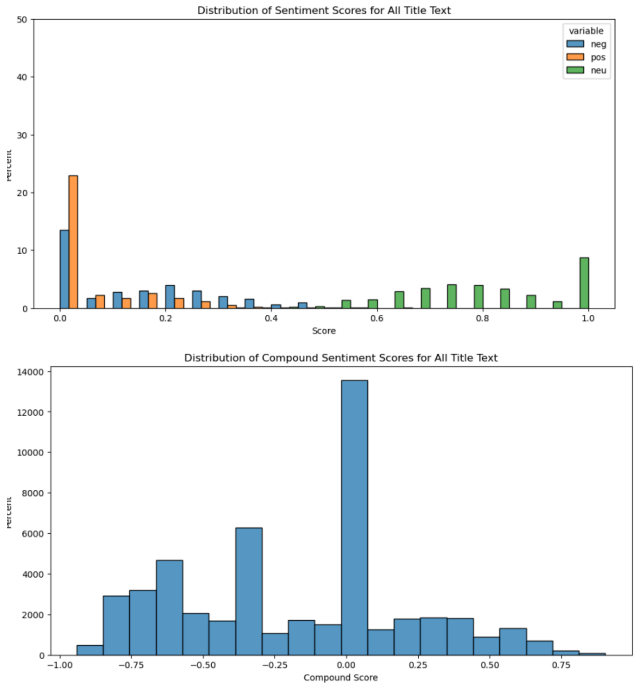

Figure 25: **Sentiment Score Distribution for Post Titles**

The overall distribution of sentiment scores is centered near neutrality, but with clear tails in both directions. This portrays a mix of objective reporting and emotionally charged headlines. The compound sentiment distribution is flatter than for comments or post text, but with more pronounced spikes around ±0.3 and ±0.6. This suggests greater variation in how events are framed.

When grouped by topic, the title sentiments show a stronger skew toward negative compound scores compared to other text fields. Some topics had markedly negative compound scores, such as Topic 9 and Topic 17, both containing keywords related to casualties, strikes, or battlefield updates. A few topics were mildly positive, such as topic 8, that were generally associated with calls to action or support.

title_topic_sentiments

Type to search

topic	pos	neu	neg	compound
0	0.028669	0.801086	0.159767	-0.207685
1	0.079306	0.786272	0.134427	-0.134768
2	0.069393	0.845281	0.085326	-0.071708
3	0.083900	0.764892	0.148587	-0.183475
4	0.056896	0.829966	0.113146	-0.094077
5	0.016990	0.846382	0.136628	-0.204171
6	0.103907	0.765990	0.130092	-0.078927
7	0.088154	0.841845	0.070001	0.023199
8	0.079053	0.905096	0.015851	0.148935
9	0.044253	0.712792	0.242951	-0.360223
10	0.076675	0.830071	0.093405	-0.021516
11	0.105547	0.763190	0.131378	-0.061612
12	0.039220	0.817027	0.143759	-0.228573
13	0.026208	0.790310	0.183549	-0.308668
14	0.083596	0.845486	0.070948	0.052524
15	0.016802	0.910417	0.072781	-0.136940
16	0.028216	0.873542	0.093177	-0.160458
17	0.068855	0.712210	0.218945	-0.373642
18	0.059905	0.711579	0.228532	-0.360287
19	0.019460	0.806563	0.173964	-0.350507

Figure 26: **Sentiment Scores for Post Title Topics**

This pattern reinforces the idea that headlines tend to emphasize conflict , danger, and urgency. From a practical standpoint, this means titles may disproportionately drive negative sentiment, regardless of the post's actual content.

5. CONCLUSION

This project provided a multifaceted analysis of Reddit discourse surrounding the Russia-Ukraine conflict, integrating social network modeling, topic extraction, and sentiment analysis to reveal how users interact, what they discuss, and how they feel.

From a network perspective, we discovered that Reddit users form a highly cohesive and structured graph. Over 95% of users belong to a single giant component, and the network exhibits strong small-world and clustering characteristics. Centrality analysis highlighted a mix of influencers: some users were extremely well-connected, while others served as strategic bridges or central hubs within tightly knit communities. Louvain modularity and k-clique detection revealed overlapping but stable community structures, and hierarchical clustering made the nested nature of these ideological or behavioral clusters clear. These structural insights suggest a landscape where information can flow efficiently, but also where echo chambers may reinforce specific viewpoints.

On the textual side, topic modeling and sentiment analysis revealed distinct patterns across comments, post content, and titles. Interestingly, comments clustered into a small number of coherent topics, while posts required a much higher topic count to reach optimal coherence—an insight that aligns with how users often respond to posts in more focused, emotionally charged language. Titles, in contrast, showed sharp sentiment skew and more polarized language, likely due to their framing function. This divergence between field types is critical to understand when designing monitoring tools or engagement strategies. We also used a two-step modeling approach to select topic numbers—starting with broad grid searches to find a sensible range and then re-running evaluations in a narrower band. This method helped us tune coherence and perplexity tradeoffs with greater precision and transparency.

Our sentiment analysis found that most discourse remained neutral or mildly polarized, with some outlier topics leaning clearly positive or negative. Posts discussing war realities, global actors, or political figures exhibited stronger sentiment skews, reflecting areas of higher emotional intensity or ideological friction.

Overall, our approach demonstrates how combining structural network metrics with semantic content modeling can yield meaningful insights into the dynamics of online communities during major geopolitical events. Our hypothesis was partially supported by the data, confirming that interaction was largely driven by a vocal minority. However, the overall sentiment of the dataset was neutral, leaning negative due to the polarizing titles and largely neutral comments.

6. FUTURE WORK

There are several natural next steps for this analysis. First, tracking the evolution of network structure and sentiment over time would reveal how crises shape discourse and connectivity patterns. Second, integrating endorsement data (e.g., upvotes, cross-posts) could quantify influence or narrative traction more precisely. Third, connecting topic clusters to specific subreddits could identify how platform structure shapes public opinion. Finally, applying bot detection or user metadata could separate organic user activity from coordinated influence operations—something increasingly relevant in modern information warfare.

7. REFERENCES

[1] Nandurkar, Tanmay, et al. ‘Sentiment Analysis Towards Russia - Ukrainian Conflict: Analysis of Comments on Reddit’. *2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP)*, 2023, pp.1–6.*IEEEXplore*,https://doi.org/10.1109/ICETET-SIP58143.2023.10151571.

[2] Guerra, Alessio, and Oktay Karakuş. ‘Sentiment Analysis for Measuring Hope and Fear from Reddit Posts during the 2022 Russo-Ukrainian Conflict’. *Frontiers in Artificial Intelligence*, vol. 6, Apr. 2023. *Frontiers*, https://doi.org/10.3389/frai.2023.1163577.

[3] Krivičić, Armin, and Sanda Martinčić-Ipšić. ‘Analyzing Sentiment of Reddit Posts for the Russia-Ukraine War’. *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, 2023, pp. 1709–14. *IEEE Xplore*, https://doi.org/10.23919/MIPRO57284.2023.10159986.