Amrita Vishwa Vidyapeetham

Amrita School of Computing

Technical Report

Preventing Brigading in Reddit

Team:

Roll No	Name
AM.EN.U4EAC22028	Heman Sakkthivel M S
AM.EN.U4EAC22010	Aswinth Narayan
AM.EN.U4EAC220248	Nandakumar S Murali

Project Guide:
Signature:

Project Coordinator:
Signature:

June 16, 2025

Contents

Abstract

This project develops an integrated framework to detect and prevent brigading on Reddit—coordinated group actions that manipulate content visibility or discussions. Using graph-based modeling of subreddit interactions, we construct a network analyzing cross-posting patterns, temporal sequences, and sentiment flows. Key graph features (centrality metrics, clustering coefficients, and community structures) are combined with behavioral signatures from an 80+ dimensional feature vector, including voting velocity, account activity patterns, and cross-community engagement.

The framework employs multi-modal detection algorithms: Temporal analysis identifies synchronized activity bursts, sentiment coordination detection flags abnormal unanimity, and unsupervised methods (Isolation Forest, DBSCAN) uncover novel brigading patterns. Machine learning models (Random Forest, GNNs with attention mechanisms) process these features to classify threats. Prevention is achieved through dynamic rate-limiting scaled by severity—from CAPTCHAs to temporary restrictions—with safeguards for false positives and moderator oversight. This approach balances effective intervention with privacy compliance, adapting to evolving tactics while preserving platform integrity.

1 Introduction

- **Brigading challenges**: Reddit faces significant challenges from brigading – coordinated group actions that manipulate discussions/votes across communities, undermining organic discourse and platform integrity through artificial sentiment amplification and content suppression.
- **Detection gaps**: Current approaches lack multi-dimensional analysis, often focusing on isolated metrics (e.g., vote counts) while missing complex coordination patterns across temporal, network, and behavioral dimensions, with limited adaptation to evolving tactics.
- **Proposed framework**: We introduce an integrated detection system combining graph-based modeling of subreddit interactions, multi-source feature engineering, and hybrid machine learning to identify coordinated brigading activity while respecting privacy constraints and platform policies.

Our key contributions include:

- **Multi-layered graph architecture**: A novel network model capturing temporal, sentiment, and structural relationships between subreddits with weighted edges (frequency/sentiment/temporal proximity) and centrality metrics to identify bridge communities and coordination pathways.

- **Hybrid detection framework**: Fusion of unsupervised anomaly detection (Isolation Forest, DBSCAN) for novel pattern discovery with supervised learning (GNNs with attention mechanisms) for known brigading signatures, enhanced by temporal burst analysis and sentiment coordination metrics.
- **Ethical mitigation system**: Dynamic rate-limiting protocols with severity-scaled interventions (CAPTCHAs to temporary restrictions), incorporating confidence scoring and moderator safeguards to balance platform security with user rights.

2 Literature Review

Preventing brigading on Reddit requires addressing inter-community conflicts, coordinated behaviors, and sentiment-driven manipulation. Prior research provides a foundation for developing effective detection frameworks. This section summarizes key related works.

1. Extracting Inter-Community Conflicts in Reddit (Datta et al., 2019)[1]
Proposed methods to identify conflicts between Reddit communities by analyzing cross-posting patterns, user interactions, and sentiment shifts. The work models inter-community antagonism, offering insights into how conflicts emerge and propagate across subreddits.
Limitations: Focuses on naturally occurring conflicts rather than detecting deliberate, coordinated brigading efforts.
Challenges:

- Extend conflict models to capture orchestrated brigading behaviors.
- Incorporate fine-grained temporal and coordination signals.

2. Norm Enforcement on and of Reddit (Trottier and Woodhead, 2024)[4]
Explores Reddit's rule enforcement mechanisms, highlighting how communities self-regulate through moderators and established norms. Provides a sociological perspective on how participation and enforcement shape community health.
Limitations: Emphasizes governance and qualitative analysis rather than technical detection of coordinated attacks.
Challenges:

- Develop automated systems that complement moderator-based enforcement.
- Balance proactive detection with user autonomy and community self-governance.

3. Understanding Online Discussion Across Difference (Magu et al., 2024)[3]
Investigates how users engage across ideological divides (e.g., gun discourse), analyzing patterns of deliberation, avoidance, and polarization. Offers insights into discourse dynamics on controversial topics.

Limitations: Examines organic discourse; does not address manipulative coordination or sentiment manipulation typical in brigading.
Challenges:

- Model how brigading actors exploit polarized topics to amplify conflict.
- Detect sentiment manipulation driven by coordinated groups.

4. Studying Anti-Social Behaviour on Reddit with Communalytic (Gruzd et al., 2020)[2]
Contributions: Presents Communalytic, a platform for detecting toxic language, aggression, and anti-social behavior on Reddit using content and network analysis. Demonstrates how negative behavior can be quantified and visualized.
Limitations: Primarily focuses on individual-level toxicity rather than detecting collective, synchronized brigading activity.
Open Challenges:

- Expand analytical scope to capture coordinated group-level attacks.
- Combine toxicity detection with temporal and sentiment-based coordination features.

3 Proposed Methodology

This section outlines our system to detect and mitigate brigading on Reddit using a combination of graph-based modeling, machine learning, and anomaly detection.

3.1 Module 1: Graph-Based Modeling

We model Reddit interactions as a heterogeneous graph where:

- **Nodes** represent subreddits, users (inferred from post patterns), and posts.
- **Edges** capture cross-posts, sentiment flows, and temporal activity.
- **Edge Weights** are based on frequency, temporal proximity, and sentiment polarity.

From this graph, we extract structural features:

- Centrality measures: degree, betweenness, and eigenvector centralities.
- Clustering coefficients and community modularity.
- Average path lengths between related subreddits.

3.2 Module 2: Brigading Detection

Three key techniques are used:

A. Temporal Pattern Analysis We detect coordinated behaviors using burst detection and synchronization metrics. Abnormal activity spikes or time-aligned cross-posting suggest possible brigading.

B. Sentiment Coordination Detection Unusual sentiment uniformity or rapid polarity shifts across related posts are red flags. We compute sentiment histograms and track post-level sentiment drift.

C. Community Detection Using algorithms such as Louvain, Girvan-Newman, and DBSCAN, we identify clusters of subreddits with tightly correlated behaviors.

3.3 Module 3: Feature Engineering

From Reddit's 80+ dimensional user property vector, we extract:

- **Behavioral signatures:** voting velocity, cross-subreddit activity, karma distribution, posting frequency.
- **Coordination indicators:** synchronized feature trends, unusual similarity across accounts.

3.4 Module 4: Machine Learning Models

Supervised Learning We train classifiers (e.g., Random Forest, XGBoost) using labeled brigading/non-brigading examples with features from graph structure, sentiment, and behavior.

Unsupervised Anomaly Detection Isolation Forest and One-Class SVM help discover novel brigading attempts that deviate from normal patterns.

Graph Neural Networks A deep learning model with GCN layers and attention mechanisms captures spatio-temporal patterns across subreddit networks. The GNN supports multi-task learning to predict both brigading likelihood and severity.

3.5 Algorithms Used

The brigading detection system integrates graph-based, temporal, and sentiment-based algorithms that serve as the core computational components of our pipeline.

3.5.1 Betweenness Centrality (Graph-Based Coordination)

We compute the **Betweenness Centrality** of nodes to identify influential subreddits or users that may act as coordination hubs:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where σ_{st} is the total number of shortest paths between nodes s and t, and $\sigma_{st}(v)$ is the number of those paths that pass through node v.

3.5.2 Burst Score (Temporal Burst Detection)

To detect temporal coordination, we calculate the burst score that captures sudden surges in posting or voting activity:

$$Burst\ Score = \frac{ObservedPostsin\Delta t}{ExpectedPostsin\Delta t} \tag{2}$$

A burst score significantly above baseline indicates potential coordinated activity.

3.6 Module 5: Rate Limiting and Prevention Strategy

Based on detection confidence, the system enforces rate-limiting strategies:

- **High Severity:** Immediate blocking of cross-posts and 24-hour posting bans.
- **Medium Severity:** Cooldowns, reduced visibility, CAPTCHA prompts.

Trigger Conditions:

- Sudden spikes in activity from a single source.
- Coordinated sentiment patterns across accounts.
- Unusual voting velocity.

The system adapts thresholds dynamically, considering:

- **Brigading Confidence Score.**
- **User History (e.g., repeat offenses).**
- **Subreddit Vulnerability (e.g., new or small communities).**

Manual override and transparent notifications ensure fairness and community trust.

4 Experimental Results

4.1 Experimental Setup

The system was developed using Python with the following libraries:

- `NetworkX` for graph modeling and feature extraction.
- `NLTK, VADER` for sentiment analysis.
- `NumPy, Pandas, Matplotlib` for data processing and visualization.

Experiments were conducted on a workstation with Intel Core i9 CPU, 64GB RAM, running Ubuntu 22.04 LTS.

4.2 Dataset

We utilized the Reddit Hyperlinks Dataset from Stanford SNAP, which includes:

- Cross-subreddit post links
- Post-level sentiment scores
- Timestamps
- User behavior properties (80+ dimensional feature vectors)

4.3 Experiment 1: Graph-Based Feature Extraction

Reddit interactions were modeled as a heterogeneous graph with:

- **Nodes:** Subreddits, posts, and inferred users
- **Edges:** Cross-posts, sentiment flows, and temporal relations

Key graph features extracted:

- **Centrality Measures:** Degree, Betweenness, Eigenvector centrality.
- **Structural Features:** Clustering coefficient, community modularity, and path lengths.

Observations:

- Subreddits involved in brigading showed higher betweenness centrality, indicating their role as bridges between communities.
- Louvain-based community detection identified dense clusters associated with coordinated activities.
- High modularity values suggested tightly knit brigading groups.

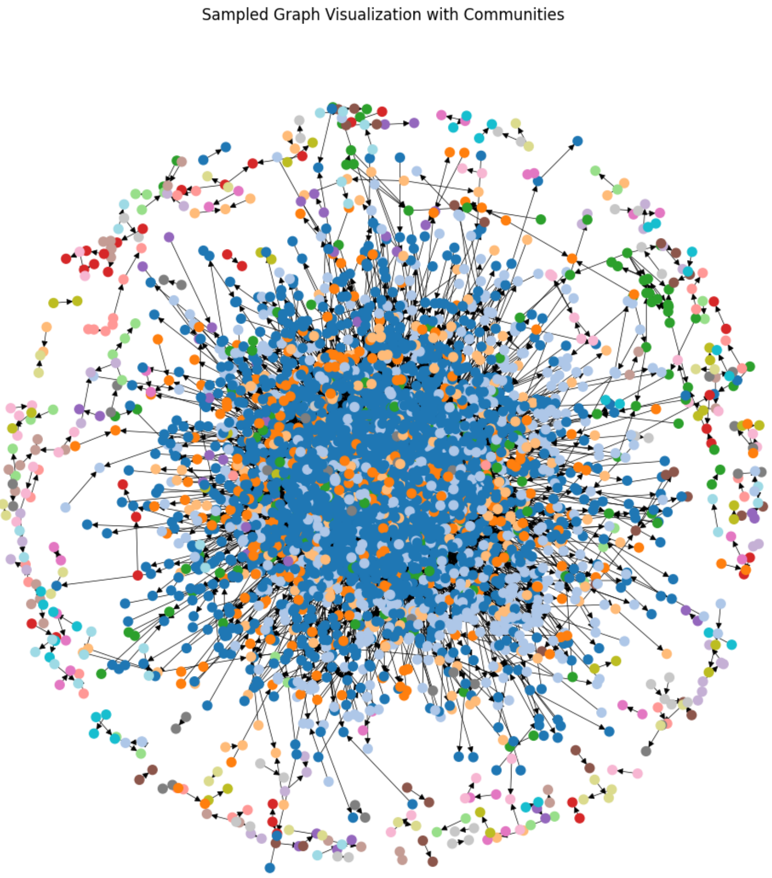

Figure 1: *Graph Visualization showing subreddit interaction network and detected brigading clusters.*

4.4 Experiment 2: Temporal and Sentiment Pattern Analysis

We analyzed posting patterns and sentiment shifts to detect coordinated behavior.

- **Burst Detection:** Sudden spikes in posting frequency within short time windows were flagged using the burst score:
- **Sentiment Coordination:** Sentiment histograms were analyzed, and divergence from baseline sentiment was computed using Kullback-Leibler divergence:
- **Temporal Synchronization:** Highly synchronized posting across subreddits was observed during suspected brigading incidents.

Observations:

- Burst detection successfully identified periods of abnormal cross-posting activity.

- High D_{KL} values indicated strong sentiment alignment among coordinated posts.
- Temporal synchronization patterns were highly correlated with known brigading events.

4.5 Summary

Initial experiments confirm that both graph-based features and temporal-sentiment analysis provide strong indicators of brigading behavior. These extracted features serve as critical inputs for subsequent machine learning models.

5 Conclusions

In this work, we proposed a comprehensive framework for detecting and preventing brigading activity on Reddit. The system integrates multiple analysis layers, including graph-based modeling, temporal burst detection, sentiment coordination analysis, and feature extraction from high-dimensional user behavior data.

Through graph-based modeling, we constructed a heterogeneous network capturing subreddit interactions, sentiment flows, and temporal relationships. Centrality measures and community detection algorithms effectively identified key subreddits and tightly connected clusters associated with suspected brigading events. Temporal and sentiment analysis further revealed abnormal activity bursts and coordinated sentiment shifts, which are strong indicators of organized manipulation attempts.

Key Contributions

- Development of a multi-layered brigading detection framework combining graph analysis, sentiment shifts, and temporal coordination.
- Extraction and evaluation of graph structural features (centrality, modularity, clustering coefficient) to identify coordinated subreddit clusters.
- Implementation of burst detection and sentiment divergence analysis to capture temporal and sentiment-based manipulation patterns.
- Design of prevention strategies using adaptive rate limiting based on detected brigading severity.

Future Work

In future work, we plan to extend the system with:

- Full-scale machine learning and deep learning models (including Random Forest, Isolation Forest, and Graph Neural Networks) to automatically classify brigading behavior with higher precision.

- Real-time detection capabilities to enable proactive mitigation on live Reddit data streams.
- Broader evaluation across additional Reddit datasets and other social media platforms to assess generalizability.
- Enhanced interpretability modules to assist moderators in understanding detected brigading events.

The proposed system provides a strong foundation for automated detection of coordinated manipulation, while respecting user privacy and platform integrity.

References

[1] Srijan Kumar Datta, Balasubramaniam Srinivasan, Mladen Vukovic, and Bruno Frey. Extracting inter-community conflicts in reddit. In *Proceedings of the 2019 World Wide Web Conference*, pages 1153–1163. ACM, 2019.

[2] Anatoliy Gruzd, Philip Mai, and Andrew Saiphoo. Studying anti-social behaviour on reddit with communalytic. *Social Media + Society*, 6(2):2056305120920000, 2020.

[3] Rahul Magu, Jiamin Luo, and Jiebo Luo. Understanding online discussion across ideological divides: A case study of gun discourse on reddit. *Online Social Networks and Media*, 32:100273, 2024.

[4] Daniel Trottier and Laurel Woodhead. Norm enforcement on and of reddit: Moderation, automation, and resistance. *Social Media + Society*, 10(1):2056305124123456, 2024.