Community Susceptibility Detection to Misinformation: News Sources & Reddit Communities on January 6th Event

WeiChun Chang, Jennifer Rozenblit
University of Texas at Austin
Electrical & Computer Engineering Department, Math Department
{weichun.chang, jrozenblit}@utexas.edu

Abstract

This project analyzes communities on social news platforms (e.g., Reddit) and their susceptibility to political misinformation. A network of news sources is constructed based on credibility, while a user embedding pipeline detects communities by examining relationships between posts and news sources. Communities are mapped onto a credibility-bias space (informed by news source interactions) to analyze their dynamics. The approach identifies communities at high risk of engaging with low-credibility or biased content and predicts their susceptibility based on affiliations. By implementing methods like ACE scoring to detect anomalous behaviors in heterogeneous networks, communities with higher susceptibility are identified. The primary contribution is the analysis of aggregated groups, media sources, and their temporal dynamics to understand how exposure and interactions influence shifts in susceptibility over time.

1. Introduction

Social news platforms like Reddit have become key hubs for information exchange, enabling users to share opinions, link to news articles, and engage in discussions. These platforms, now integral to news consumption, host diverse viewpoints but also facilitate the spread of unverified or biased information. For instance, Reddit's r/politics, the largest political news discussion community, contains a substantial amount of unverifiable content amplified by recommendation algorithms. This fosters echo chambers, reinforcing pre-existing beliefs while excluding opposing perspectives, and spreading biased narratives[1].

Echo chambers become especially problematic during politically charged events, such as the January 6th, 2021, U.S. Capitol attack. The fragmented media environment during such events often increases susceptibility to misleading narratives, shaping public perception and collective beliefs.

Research has addressed misinformation using deep learning to detect unreliable sources[2, 3], large language models for similar tasks[4], and moderation strategies targeting users spreading unreliable content[5]. However, comprehensive analyses combining source credibility with sharer biases remain scarce.

This paper bridges this gap by analyzing user credibility, media source bias, and reliability over time. It examines community susceptibility to unreliable content, leveraging datasets from the Global Database of Events, Language, and Tone (GDELT) for major news narratives and Reddit posts for community reactions. A credibility network evaluates media source reliability with models like SentenceBERT[6] and ChatGPT-4 to detect outlet conflicts[7].

We propose a comment-based user embedding methodology to represent users in a latent space, analyzing their interactions with biased content. This approach explores news source interaction and bias distribution in communities, addressing:

1. How do media sources differ in credibility on key political topics?
2. How do Reddit communities vary in susceptibility to misinformation regarding credibility and bias?
3. Can engagement patterns predict misinformation susceptibility in communities?

Our findings reveal mechanisms underlying misinformation spread and identify at-risk communities, informing strategies to combat its societal impacts.

2. Background & Work

2.1 Sentence Embedding and Semantic Analysis

Sentence embedding encodes textual data into numerical vectors, preserving syntactic and semantic relationships for tasks like clustering and similarity analysis. BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018) [8] introduced a bidirectional architecture for deep linguistic representation. RoBERTa [9] improved BERT by removing next-sentence prediction and pretraining on larger datasets. SentenceBERT (SBERT) by Reimers and Gurevych (2019) [6] optimized sentence-level tasks with a Siamese architecture for efficient pairwise similarity computation. DistilBERT[10] and T5 further enhance sentence embedding by balancing efficiency and semantic understanding.

These advancements enable sophisticated semantic analysis tools for large-scale text data, such as social media interactions.

2.2 Heterogeneous Multiplex Network

A Heterogeneous Multiplex Network (HMN) is defined as a quintuple $G = (V, E, L, T, \mathcal{R})$ where: V is the set of nodes, $E \subseteq ((V \times L) \times (V \times L))$ is the set of edges connecting vertices across layers L, and$T = \{T_V, T_E\}$ is the set of types, namely of of vertex types (T_V) and edge types (T_E). Note these types are equipped with relations from $\mathcal{R}$, which is the set of functions mapping between vertices, edges, and types:

The mapping functions are defined as $R_{VT} : V \to T_V$ and $R_{ET} : E \to T_E$. A multiplex network maps an equivalent node to itself from one layer to the next.

2.3 Anomaly Detection

Dynamic graph anomaly detection addresses structural and temporal changes in networks. Distance-based methods, such as Graph Edit Distance and Hamming distance, capture structural irregularities but are computationally intensive for large graphs. Compression-based approaches, like GraphScope, detect anomalies by encoding graphs compactly using Minimum Description Length principles. Decomposition techniques, such as Compact Matrix Decomposition (CMD), identify anomalies through reconstruction errors, although they may miss localized patterns[11]. Community-based methods, like those by Aggarwal et al.[12], focus on changes within graph clusters, offering scalability for massive networks.

Probabilistic models use Bayesian frameworks to model normal graph behaviors, while window-based techniques, such as PLADS, maintain anomaly detection within temporal partitions to ensure scalability. Centrality-based methods, particularly the ACE score[13], provide a balanced approach by integrating structural and temporal anomalies, aligning with the challenges of evolving networks.

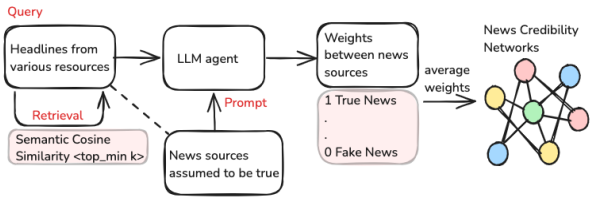

Figure 1: Credibility Network Architecture

Our approach adopts the ACE methodology, leveraging centrality metrics for detecting deviations in dynamic graphs, enabling efficient and precise anomaly detection in community behaviors.

3. Methodology & Experiment

This research integrates two networks: one assessing news source credibility and another analyzing Reddit community reactions. The first evaluates the consistency and agreement of media sources, while the second studies community engagement over time, focusing on exposure to conflicting narratives. The methodology involves the following steps:

3.1 Key Steps

1. **Headlines sentence embeddings** News articles are gathered from diverse sources across the political spectrum using fixed keywords and date ranges for consistency. Relevance and diversity are ensured through threshold-based filtering. We first extract news headlines by searching for specific keywords, such as “Trump,” “Capitol,” and “riot,” that are representative of key political events. These keywords guide the selection of relevant articles for analysis. A neutral query text, such as the example provided below, is then fed into the model for the process of sentence embedding. Example Query: ”On January 6, 2021, the United States Capitol Building in Washington, D.C., was attacked by a mob of supporters of then...” Events such as the January 6th Capitol attack are analyzed with semantic similarity models (e.g., BERT) to extract representative headlines.
2. **LLM Analysis of Source Credibility:** Large language models (LLMs) assess media credibility by comparing narratives across sources. For instance, CNN headlines are benchmarked against Fox News to identify biases and factual consis-

tency. Using pre-trained models like sentence-BERT (all-MiniLM-L12-v2), query text is embedded and semantic similarity calculated between the query and headlines. The top k headlines from each source (cosine similarity ≥ 0.05, up to 30 headlines) are analyzed with ChatGPT. The LLM is prompted: "Assume all information is factual; judge other sources. Score 0 (fake news) to 1 (true news) with a 10-word explanation." This allows the LLM to evaluate headline alignment with the reference (e.g., CNN) and score factual accuracy. Repeating this for all sources generates a credibility network, where nodes represent media outlets, and weighted edges reflect coverage alignment or conflict, based on averaged credibility scores.

3. **Multilayer Network Construction** A multilayer network is constructed to capture temporal dynamics. We have two types of vertices: news sources and reddit posts from r/politics, r/democrat, r/republican, r/liberal, r/conservative, and r/libertarian that contain news source references. Similarly, we have two types of edges: post to news source edges, representing the relevance of the post in the community that references that news source (where relevanc e is defined to be the fraction of upvotes the post recieved relative to all upvotes in that community/thread) and news-news edges which are from the previously mentioned news source credibility network. Each layer represents a different date (January 6th, 7th, and 8th). Daily credibility scores for each news source are tracked, enabling an analysis of how credibility and community interaction evolves over time.
4. **Anomaly Detection with ACE Score:** Anomaly detection is performed using the ACE score in the heterogeneous dynamic network. This approach aligns with a three-layer architecture:
    - **Input Layer:** Nodes (Reddit posts and media sources) and edges (weighted by relevance or truth) form the network structure. Temporal decomposition enables day-wise analysis, capturing evolving dynamics.
    - **Processing Layer:** For each Reddit node, the ACE score is computed as:

$$ACE(v) = \frac{\sum_{u \in N(v)} w(u, v)}{\deg(v)}$$

      where v is a Reddit node, $N(v)$ are neighboring media nodes, and $w(u, v)$ is the edge weight. The anomaly detection is performed using the Isolation Forest algorithm, which is configured to identify the top 5 percent of data points as potential anomalies. The algorithm is applied to the scaled node features, and the results are used to identify anomalous Reddit nodes. Results are grouped to identify trends in susceptibility across ideological and temporal layers. Outliers are flagged based on ACE score distributions, identifying potential misinformation hotspots or isolated nodes.
    - **Output Layer:** Ranked lists of posts by ACE score, aggregated insights by community and time, and flagged anomalies are produced. These outputs address project goals, such as identifying influential narratives and vulnerable communities.
5. **Aggregation by Community and Temporal Evolution Analysis:** ACE scores are aggregated at the community (subreddit) and temporal (day) levels. This analysis uncovers patterns of susceptibility to influential or polarizing narratives. Communities (e.g., `conservative`, `liberal`) and specific days (e.g., `day2`) with anomalous ACE scores are highlighted as potential misinformation hotspots or outliers, signaling amplification or isolation issues.
6. **Outcome and Analysis:** The multilayer credibility network facilitates an analysis of source credibility and community interactions over time. By examining how Reddit communities engage with news sources, particularly those with potential biases or low credibility, we identify communities that are vulnerable to misinformation and assess the impact of narrative shifts on media trust.

This methodology integrates media credibility analysis with community dynamics, offering a comprehensive framework for understanding the dissemination of misinformation and the influence of user engagement on media trust.

3.2 Assumptions

The approach relies on the following assumptions:

1. **Accuracy of LLM Judgments:** LLMs are assumed to provide reliable assessments of media biases, supplemented by human review to address limitations such as training data biases.
2. **Representativeness of Subreddit Data:** Selected subreddits are assumed to reflect broader online engagement with political news, though we account for potential cultural or ideological biases in individual communities.

3. **Temporal Stability of Community Behavior:** Historical patterns of community engagement with misinformation are presumed indicative of future behaviors, allowing predictive modeling based on past interactions.
4. **Definition of Credibility and Network Construction:** Credibility is defined as a composite measure including reputation, factual consistency, and citation diversity. The credibility network uses these assessments, capturing temporal shifts in opinions and identifying communities prone to misinformation.

4 Experiment Result

4.1 Results

The similarity score using sentence embedding illustrate the ideological alignment of news sources through a credibility network constructed using LLM evaluations. For the January 6th Capitol riots, scores indicate how closely other outlets align with the ground truth's narrative, with alignment values ranging from 0 (no alignment) to 1 (full alignment). Headlines attributing direct responsibility to Trump scored higher, reflecting closer alignment with CNN, while neutral or unrelated headlines, such as Covid, scored lower. These findings highlight ideological variance across headlines and demonstrate that media bias is topic-specific, varying even within the output of individual outlets.

Table 1: Example: Cosine Similarity of CNN Headlines on 2024-01-06

Headline	Cosine Similarity
Democrats are on the verge of taking the Senate	0.2790
Multiple officers injured in the mob violence	0.2189
Congress has the final vote in the 2020 election today. Here how it will work.	0.2114
Vice President Pence calls for rioters to leave the Capitol building	0.2016
Tracking the electoral vote count in Congress	0.1956
NYSE bans Chinese telecom stocks in second about-face of the week	0.1749
Electoral vote count in Congress: A step-by-step guide	0.1593
Here where the vote count stands in Georgia	0.1381
WHO Covid team blocked from entering China to study origins of coronavirus	0.1316

The credibility network, constructed using ChatGPT-assessed factuality scores, comprises nodes representing news outlets, color-coded by political leaning: red (right-leaning), blue (left-leaning), and green (neutral). Reuters stands out as the most credible source, attributed to its factual and concise reporting that avoids emotionally charged or partisan language.

On the other hand, in the heterogeneous network, subreddit nodes connect to news resources, reflecting community interactions. We find that CNN emerges as a central, highly connected node, consistently ranking highest in Degree, Centrality, and PageRank metrics, underscoring its role as a key information hub. By contrast, peripheral nodes like The Federalist exhibit lower influence, indicative of their more isolated positions within the network.

Table 2: Network Edge Weights between Nodes

Source	Target	Weight	Source	Target	Weight
cnn	reuters	red!250.9273	foxnews	reuters	red!250.7591
cnn	theepochtimes	0.3067	foxnews	theepochtimes	0.543
cnn	nytimes	0.7	foxnews	nytimes	0.6693
cnn	foxnews	0.545	foxnews	cnn	0.5038
cnn	thefederalist	0.2	foxnews	thefederalist	0.16
cnn	washingtontimes	0.3983	foxnews	washingtontimes	0.535
cnn	latimes	0.5133	foxnews	latimes	0.5567
cnn	washingtonpost	0.6317	foxnews	washingtonpost	0.3967

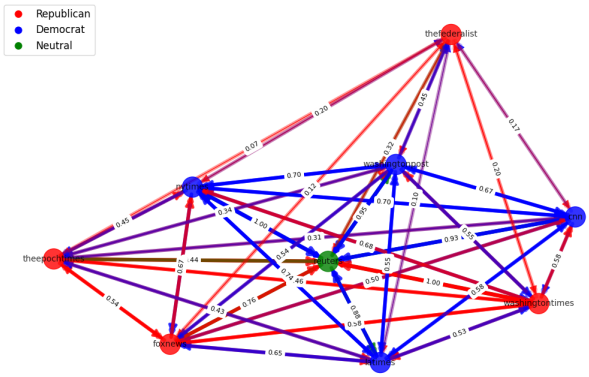

Figure 2: Credibility Network of news sources on January 6

Table 3: Expanded Network Statistics for All Layers

Layer	Nodes	Edges	Avg Degree	Density	Diam	Avg Path Len	Modularity
Jan 6	270	262	1.9407	0.0036	4	2.2352	0.6834
Jan 7	354	346	1.9548	0.0028	4	2.8240	0.6820
Jan 8	324	315	1.9444	0.0030	2	1.9759	0.8460

Temporal analysis reveals trends in community interactions and media influence. On January 7th (Day 2), central nodes like CNN and Reuters experience a connectivity peak, correlating with increased engagement following the Capitol riots. By Day 3, their connectivity slightly declines as discussions shift. Peripheral nodes, meanwhile, exhibit reduced engagement over time, suggesting diminished relevance.

Table 3 and Figure 3 highlight the shifting narrative dynamics across Reddit communities, revealing varying susceptibility patterns.

Show in table 4, conservative subreddits, such as `conservative_jan_7`, consistently exhibit the highest ACE scores, suggesting strong ties to influential narratives. In contrast, Democrat and libertarian communities, while engaged, show lower ACE scores, indicating moderate susceptibility to misinformation. Temporal trends indicate Day 2 as the peak for narrative ampli-

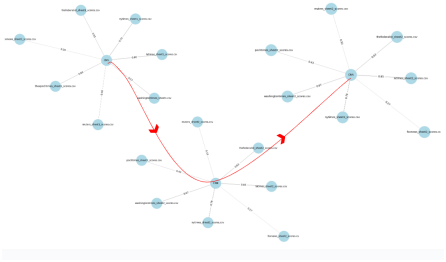

Figure 3: News Network Time Shift

Community and Network Graph	ACE Score
conservative_jan_7', 'day2_network_graph'	0.4273
democrats_jan_7', 'day2_network_graph'	0.2608
republican_jan_7', 'day2_network_graph'	0.1493
politics_jan_6', 'day1_network_graph'	0.0899
libertarian_jan_7', 'day2_network_graph'	0.0839
libertarian_jan_8', 'day3_network_graph'	0.0765
politics_jan_8', 'day3_network_graph'	0.0662
conservative_jan_8', 'day3_network_graph'	0.0608
liberal_jan_6', 'day1_network_graph'	0.0565
politics_jan_7', 'day2_network_graph'	0.0559
conservative_jan_6', 'day1_network_graph'	0.0312
democrats_jan_8', 'day3_network_graph'	0.0276
republican_jan_7', 'day2_network_graph'	0.0149
liberal_jan_7', 'day2_network_graph'	0.0147
republican_jan_8', 'day3_network_graph'	0.0146
democrats_jan_6', 'day1_network_graph'	0.0136
conservative_jan_7', 'day2_network_graph'	0.0131
politics_jan_6', 'day1_network_graph'	0.0089
politics_jan_8', 'day3_network_graph'	0.0067
democrats_jan_8', 'day3_network_graph'	0.0060

Table 4: Aggregate ACE scores for various communities and their corresponding network graphs.

fication, Day 1 as chaotic real-time responses, and Day 3 as a period of declining activity.

High-ACE posts, primarily from Day 2, focus on themes like Capitol rioter accountability, media bias, and platform actions. Anomalous posts from Day 1, covering both the riots and unrelated events such as Georgia's runoff election, underscore early narrative variability. Conservative communities dominate high-ACE and anomaly subsets, reinforcing their role as echo chambers, although Democrat and libertarian communities also show engagement with these narratives, suggesting a broader ideological susceptibility.

These findings emphasize the importance of targeted interventions, such as educating moderators in conservative subreddits, fostering cross-ideological discussions, and enhancing media transparency to reduce misinformation and promote healthier online discourse.

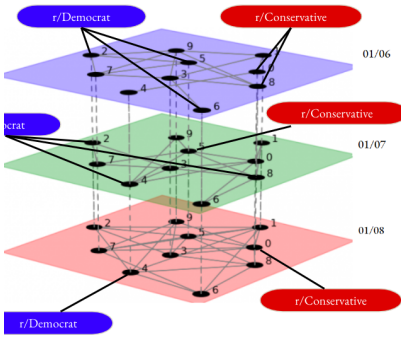

Figure 4: Layered Network

Figure 5: Comparison of News Network Time Shift and Layered Network

4.2 Challenges

Major challenges include:

- Data Scarcity: The third day of analysis (Jan 8th) presents a significant challenge due to the limited availability of data, which may affect the robustness of conclusions drawn for that period.
- Heterogeneous Network: The network consists of nodes representing both news sources and Reddit posts, which differ significantly in structure, content, and interactions. This heterogeneity complicates the analysis, as traditional network metrics may not apply uniformly across both types of nodes.
- Distinguishing Noise from True Anomalies: High ACE scores may result from natural engagement or viral neutral content, making it difficult to differentiate between legitimate activity and amplification of misinformation. This increases the risk of false positives or negatives, especially when subtle misinformation tactics evade detection.

5 Conclusion

Our findings reveal the role of conservative-leaning sources, such as Fox News, in amplifying specific narratives, while neutral outlets like Reuters demonstrated greater factual consistency. We also observed differing patterns of media susceptibility across online communities, notably within conservative subreddits.

This work's impact lies in quantitatively assessing media credibility and detecting biases through a scalable, automated framework for analyzing misinformation dynamics. It provides valuable insights to promote

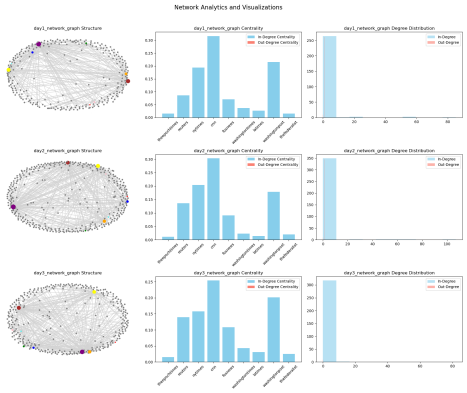

Figure 6: Heterogeneous network and characteristics

media transparency and enhance media literacy, fostering healthier public discourse.

Future research could expand this framework by integrating random walks on heterogeneous networks for comprehensive credibility assessment and improving real-time anomaly detection. Analyzing a broader range of ideologies and platforms, alongside developing tailored interventions for at-risk communities, could further mitigate misinformation and reduce polarization, strengthening societal trust in media.

6 Contribution

WeiChun drafted the project paper, documentation, literature review, and constructed the credibility news network. She wrote the experiment code and provided result interpretations. Jennifer recorded the M1 and M2 videos, built the subreddit posts network, and calculated ACE scores, contributing to result interpretation. Together, the team collaborated closely with mentor Arash Amini on the experimental design and discussions.

References

[1] G. De Francisci Morales, C. Monti, and M. Starnini. No echo in the chambers of political interactions on reddit. *Scientific Reports*, 11(1):2818, 2021.

[2] K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu. The role of user profiles for fake news detection. In *ASONAM*, pages 436–439. ACM, 2020.

[3] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint*, 2019.

[4] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, and P. Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *arXiv preprint*, 2023.

[5] F. Sakketou, J. Plepi, R. Cervero, H. J. Geiss, P. Rosso, and L. Flek. FACTOID: A new dataset for identifying misinformation spreaders and political bias. In *Language Resources and Evaluation Conference*, pages 3231–3241. European Language Resources Association, 2022.

[6] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*. ACL, 2019.

[7] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319v2*, 2024.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 2019.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint*, 2019.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[11] J Sun, Y Xie, H Zhang, and C Faloutsos. Less is more: Sparse graph mining with compact matrix decomposition. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 1(1):6–22, 2008.

[12] CC Aggarwal, Y Zhao, and PS Yu. Outlier detection in graph streams. In *2011 IEEE 27th International Conference on Data Engineering (ICDE)*, pages 399–409. IEEE, 2011.

[13] Asep Maulana and Martin Atzmueller. Centrality-based anomaly detection on multi-layer networks using many-objective optimization. In *2020 7th International Conference on Control, Decision and Information Technologies (CoDIT)*, volume 1, pages 633–638. IEEE, 2020.