Belief Is All You Need: Modeling Narrative Archetypes in Conspiratorial Discourse

Anonymous ACL submission

Abstract

Conspiratorial discourse is increasingly embedded in digital communication, yet its structure remains poorly understood. Analyzing Singapore-based Telegram groups, we show that conspiratorial content is integrated into everyday discussion rather than isolated echo chambers. We propose a two-stage framework: (1) RoBERTa-large classifies messages as conspiratorial or non-conspiratorial (F1 = 0.866 on 2,000 expert-annotated messages); (2) a signed belief graph models belief alignment via signed, similarity-weighted edges and is learned using a Signed Belief Graph Neural Network (SiBeGNN) with Sign-Disentanglement Loss to separate ideological alignment from narrative style. Hierarchical clustering of 553,648 messages reveals seven narrative archetypes: General Legal Topics, Medical Concerns, Media Discussions, Banking and Finance, Contradictions in Authority, Group Moderation, and General Discussions. SiBeGNN substantially outperforms standard methods (cDBI = 8.38 vs. 13.60–67.27), with 88% inter-rater validation. Findings show conspiratorial discourse permeates mundane domains such as finance, law, and daily life, challenging assumptions of isolated online radicalization. The framework advances belief-aware discourse modeling for low-moderation platforms and informs stance detection, political discourse analysis, and content moderation policy.

1 Introduction

The Web has transformed information circulation, enabling rapid narrative diffusion while introducing challenges for information integrity and public trust. Conspiratorial content, narratives that attribute major social, political, or health events to covert groups with hidden, malevolent intent, exemplify these challenges (Zeng et al., 2022). Such narratives oppose official or mainstream explanations, alleging deliberate deception by institutions, governments, or corporations. Prior research shows that conspiratorial discourse undermines trust in science and governance, intensifies polarization, and accelerates misinformation during crises (Douglas et al., 2017; Uscinski and Parent, 2014; Zollo et al., 2017). Understanding how these narratives emerge, circulate, and embed within digital ecosystems is therefore a key concern for web science and computational social science.

Digital platforms differ markedly in how they afford conspiratorial discourse. Telegram has become a prominent venue, particularly during crises, as users seek alternative or counter-mainstream perspectives. Its hybrid design, combining private messaging, large public groups, and frictionless forwarding, facilitates the spread of conspiratorial and agenda-driven content (Urman and Katz, 2020). During the COVID-19 pandemic, Telegram gained traction in Singapore as a space for discussing health policies, personal experiences, and political opinions, often expressing skepticism toward institutional authority (Ng and Loke, 2020). These dynamics highlight broader web science questions about how platform design shapes discourse patterns and belief formation.

Singapore provides a distinctive socio-technical context for examining conspiratorial discourse online. With high social media penetration and strong state investment in information governance, its media ecosystem reflects tensions between centralized regulation and decentralized, low-moderation platforms such as Telegram (van der Linden et al., 2023; Tandoc et al., 2021). Prior work shows that these spaces can function as echo chambers for anti-vaccine sentiment, political discontent, and conspiratorial worldviews, particularly during public health crises (Walther and McCoy, 2021; Nainani et al., 2022). However, empirical research on how conspiratorial narratives are structured, disseminated, and reinforced within closed or semi-public messaging systems in Southeast Asia remains limited (Alvern Cueco Ligo et al., 2025; Goyal et al.,

2025). This gap is consequential given the region's high digital connectivity and political diversity, where platforms like Telegram mediate both social mobilization and the diffusion of narratives shaping public opinion, health behavior, and political stability.

Recent large-scale web studies have begun mapping Telegram's conspiratorial ecosystems. The *Schwurbelarchiv* project documents multimodal conspiratorial content across German-language Telegram networks (Angermaier et al., 2025), while the *TeleScope* dataset captures longitudinal diffusion across millions of messages and channels (Gangopadhyay et al., 2025). Together, these efforts underscore the need for geographically and culturally grounded analyses that account for local contexts of meaning-making, policy discourse, and narrative co-construction.

In this study, we analyze Singapore-based Telegram groups to characterize conspiratorial discourse in a low-moderation environment. We examine not only the linguistic properties of conspiratorial messages but also their propagation and interaction within public group structures. To do so, we organize messages into *narrative archetypes*; recurring thematic patterns that reflect how meaning, intention, and stance are structured. A narrative archetype captures the narrative function a message performs (e.g., framing events, asserting causality, or reinforcing group identity), rather than its surface linguistic form. This abstraction enables analysis of the deeper narrative mechanisms through which conspiratorial discourse is constructed, shared, and sustained.

We ask: How is conspiratorial discourse structured within Singapore-based Telegram groups, and what narrative archetypes emerge from message-level patterns?

2 Related Work

Research on conspiratorial discourse spans psychology, communication, and computational social science. Foundational work defines conspiracy theories as belief systems that attribute major events to secret, malevolent groups, often driven by uncertainty, mistrust, or identity threat (Douglas et al., 2017; Uscinski and Parent, 2014). These narratives persist due to cognitive biases, affective polarization, and resistance to correction (Abdou et al., 2021; Tandoc et al., 2021). However, much of this work underemphasizes how platform-level affordances shape collective conspiratorial dynamics (Chen et al., 2023a; Goyal et al., 2023).

Telegram as a Platform for Conspiratorial Discourse: Telegram has become a key venue for conspiratorial and extremist communication due to its hybrid public-private structure, large groups, and minimal moderation (Urman and Katz, 2020; Skarzauskiene et al., 2025). Forwarding and channel interconnectivity facilitate narrative diffusion and semi-private echo chambers (Nainani et al., 2022; Chen et al., 2023b). Large-scale corpora like *Schwurbelarchiv* (Angermaier et al., 2025) and *TeleScope* (Gangopadhyay et al., 2025) enable multimodal, longitudinal analysis, but research remains mostly Euro-American, overlooking regional linguistic and cultural contexts (Ng and Loke, 2020; Goyal et al., 2025). We address this gap by focusing on Singapore-based Telegram groups.

Computational Detection of Conspiratorial Content: Transformer-based models like BERT, RoBERTa, DeBERTa, and newer systems such as LLaMA and SEA-LION have advanced automated detection (Urman and Katz, 2020; Skarzauskiene et al., 2025). Though effective on benchmarks, performance drops under domain shift and multilingual noise (Walther and McCoy, 2021; Alvern Cueco Ligo et al., 2025). Existing approaches treat detection as binary classification, ignoring diverse communicative patterns and narrative structures. We go beyond this by identifying and characterizing distinct narrative archetypes.

Graph Neural Networks for Social Discourse: Graph-based methods naturally capture relational structures in online discourse. Signed graph neural networks model positive and negative edges to represent agreement and opposition, in line with structural balance theory (Derr et al., 2018; Zhang et al., 2021, 2024). Prior applications focus on explicit social ties or simple sentiment, rarely addressing complex belief alignment and ideological conflict in text networks. Most also lack mechanisms to disentangle orthogonal dimensions, such as belief versus stylistic expression. SiBeGNN addresses this with a novel Sign-Disentanglement Loss.

Disentangled Representation Learning: Models such as DisenGCN (Ma et al., 2019a) and DiGGR (Hu et al., 2024) demonstrate how orthogonal subspaces can isolate distinct generative processes in graph-structured data. However, they have not been adapted to signed graphs or belief-oriented text

networks. We bridge this gap by integrating disentangled representation learning with signed belief graphs, enabling simultaneous capture of ideological alignment and communicative style.

Clustering and Narrative Archetype Discovery: Clustering methods reveal latent communities and thematic patterns from text-network embeddings (Li et al., 2021), including hierarchical clustering, HDBScan, Gaussian Mixtures, and topic models. While these capture semantic similarity, they do not model belief polarity or antagonism, limiting their ability to uncover ideological structures in conspiratorial discourse, where messages may be thematically diverse but share skepticism (Uscinski and Parent, 2014). No prior work systematically clusters conspiratorial content to identify narrative archetypes based on both semantic content and belief alignment; existing studies rely on binary classification/qualitative thematic analysis.

3 Data

We use two datasets corresponding to our two-stage pipeline: an annotated subset for conspiratorial classification (Stage 1) and a large-scale corpus for narrative archetype discovery (Stage 2).

Stage 1: Annotated Training Set. We randomly sampled 2,000 Telegram messages for expert annotation. Two experts with over five peer-reviewed publications each labeled messages using the definition from Diab et al. (2024): *"A conspiracy theory is a narrative accusing agent(s) of specific actions serving secretive and malevolent objectives."* Inter-annotator agreement was 85%, with a third expert resolving disagreements. The balanced dataset (1,000 conspiratorial; 1,000 non-conspiratorial) was used to fine-tune a RoBERTa-based classifier.

Stage 2: Full Corpus. We collected complete histories from six Singapore-based Telegram groups with diverse themes (Table 1), yielding **553,648 messages** and **24,270,653 words** (mean = 43.8 words, SD = 70). The trained classifier labels all messages, enabling signed belief graph construction with edges encoding textual similarity and belief alignment. SiBeGNN learns disentangled embeddings, which are hierarchically clustered to identify narrative archetypes.

Dataset	Words	Messages	Dates (From – To)
1M65	11,213,414	524,524	19-12-08 – 25-02-04
Chill Corner	99,291	12,580	25-04-15 – 25-05-14
Healing The Divide	10,687,504	3,919	21-08-11 – 25-01-20
Mile Lion	194,767	10,900	25-04-15 – 25-05-14
SG Corona Freedom Lounge	1,623,387	829	21-06-14 – 25-01-19
SG Covid Infection Survivor	452,290	896	21-07-31 – 25-01-19

Table 1: **Overview of Telegram Groups Analyzed.** Summary of word counts, message volumes, and temporal coverage for each dataset.

4 Methods

We use a two-stage approach: (1) fine-tune RoBERTa-large to classify messages as conspiratorial or non-conspiratorial; (2) build a signed belief graph with edge signs for belief alignment and weights for similarity, then apply a Signed Belief Graph Neural Network (SiBeGNN) to learn embeddings disentangling belief polarity from narrative style, which are hierarchically clustered into seven narrative archetypes.

4.1 Stage 1: Conspiratorial Content Classification

The first stage of our methodology involves training a binary classifier to distinguish conspiratorial from non-conspiratorial discourse. To identify the optimal language model for conspiratorial content classification, we conducted a systematic comparison of nine transformer-based architectures, selected based on their documented performance on discourse classification tasks and regional linguistic applicability. The evaluated models include: RoBERTa-large (Liu et al., 2019), Gemini 2.0 Flash (Google DeepMind, 2024), LLaMA 3.2 3B (Meta AI, 2024), RoBERTa-base (Liu et al., 2019), DeBERTa-base (He et al., 2021), BERT-large-uncased (Devlin et al., 2019), DistilBERT-base-uncased (Sanh et al., 2019), BERT-base-uncased (Devlin et al., 2019), and aisingapore/SEA-LION-v1-7B (AI Singapore, 2024). SEA-LION-v1-7B is a multilingual large language model specifically developed for Southeast Asian languages by AI Singapore, included to assess the potential advantages of region-specific pretraining. All models were fine-tuned on the balanced 2,000-message manually annotated dataset using identical preprocessing and tokenization pipelines to ensure comparability. Annotation was performed by expert co-authors of this paper rather than external participants. Annotators were provided with a written definition of conspiratorial discourse from (Diab et al., 2024), along with examples and counterexamples discussed during a calibration round prior to annotation.

Training was conducted for three epochs with a learning rate of 2×10^{-5}, batch size of 16, and

evaluation conducted after each epoch. Model performance was assessed using accuracy, precision, recall, and F1-score on a held-out test set. The best classifier is then applied to the full corpus of 553,648 messages to generate binary belief labels (conspiratorial vs. non-conspiratorial). An example of both conspiratorial and non-conspiratorial message is shown in Table A1

4.2 Stage 2(a): Signed Belief Graph Neural Networks

Using the binary conspiratorial labels, we construct a graph capturing semantic similarity and belief polarity. We then apply our Signed Belief Graph Neural Network (SiBeGNN) to learn embeddings for messages.

4.2.1 Graph Construction and Representation

Telegram messages were cleaned and encoded using a fine-tuned RoBERTa-large to obtain contextual semantic embeddings, which were augmented with four discourse-level features to capture how beliefs are expressed in addition to message content. Epistemic modality measures expressed certainty or uncertainty through hedging and certainty markers, capturing degrees of belief (Egan and Weatherson, 2011; Markkanen and Schröder, 1997). Agency captures active versus passive framing by comparing active and passive constructions, reflecting speaker positioning in discourse. Sentiment polarity is derived from a transformer-based classifier and scored continuously in [–1, 1] to represent positive, negative, or neutral affect (Chriqui and Yahav, 2022). Emotion spectra encode probabilities of joy, anger, fear, and sadness using a pretrained emotion model, producing a multidimensional affective representation (Araque et al., 2019). These seven discourse features are concatenated with RoBERTa embeddings, mean-centered, and L2-normalized to form enriched discourse-belief vectors that separate belief content from expressive style.

Using these representations, we construct a signed graph $G = (V, E, W, S)$ where nodes correspond to messages, edge weights $W_{ij} \in [0, 1]$ represent cosine similarity between discourse-belief embeddings, and edge signs $S_{ij} \in +1, -1$ encode belief alignment, positive for messages sharing the same conspiratorial label and negative for opposing labels. To ensure sparsity, edges are retained only when similarity exceeds 0.5, with positive edges above $\mu + 0.5\sigma$ and negative edges below $\mu - 0.5\sigma$, where μ and σ denote the mean and standard deviation of similarity. The resulting signed belief graph captures the socio-semantic polarity of discourse, explicitly modeling agreement and antagonism and distinguishing cooperative clusters that reinforce conspiratorial narratives from adversarial clusters that contest or debunk them (Awal et al., 2022; Chin et al., 2024).

4.2.2 Disentangled Sign-Aware Graph Neural Network

Building on signed network representation learning (Kumar et al., 2016) and disentangled graph embeddings (Ma et al., 2019b), SiBeGNN learns separate representations for belief polarity and narrative style from the signed graph. Each node v_i is mapped into two orthogonal subspaces: a belief subspace $z_{b,i} \in \mathbb{R}^{d_b}$ encoding ideological alignment, and a persona subspace $z_{p,i} \in \mathbb{R}^{d_p}$ capturing stylistic and behavioral traits independent of belief. The final embedding is the concatenation $z_i = [z_{b,i}; z_{p,i}] \in \mathbb{R}^{d_b+d_p}$.

Architecture: The model begins with learnable node embeddings $X \in \mathbb{R}^{N \times d_{\text{in}}}$, processed through two sign-specific graph convolutional layers, one for positive edges E^+ and one for negative edges E^-. Message passing is defined as:

$$\begin{aligned} h^+ &= \text{ReLU}(\text{GCNConv}^+(X, E^+)), \\ h^- &= \text{ReLU}(\text{GCNConv}^-(X, E^-)), \end{aligned} \tag{1}$$

where GCNConv denotes graph convolutional operations. Hidden states are concatenated and projected:

$$h = \text{ReLU}(W_p[h^+; h^-]), \tag{2}$$

where W_p is a learnable weight matrix. Finally, belief and persona embeddings emerge through separate linear projections, both L2-normalized for stability:

$$z_b = W_b h, \quad z_p = W'_p h, \tag{3}$$

This architecture performs sign-aware message passing while explicitly disentangling ideological and behavioral signals.

Sign-Disentanglement Loss: Training uses the Adam optimizer with weight decay, selecting the checkpoint with lowest total loss. On convergence, SiBeGNN yields two orthogonal embeddings, $z_{b,i} i = 1^N$ for belief polarity and $zp, i_{i=1}^N$ for narrative persona, used for hierarchical clustering in Stage 2(b). A composite loss optimizes SiBeGNN through four objectives enforcing structural fidelity, semantic grounding, and disentangled

representation. An ablation of performance of cluster quality with respect to different components of the loss is shown in Table A3.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{sign}}\mathcal{L}_{\text{sign}} + \lambda_{\text{belief}}\mathcal{L}_{\text{belief}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} \tag{4}$$

(1) Reconstruction Loss preserves signed structural relationships. Predicted adjacency is:

$$\hat{A}_{ij} = \sigma(z_i^\top z_j), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function. Target adjacency $A_{\text{target}} \in [0,1]^{N\times N}$ maps positive edges to 1, negative edges to 0, and unobserved edges to 0.5:

$$A_{\text{target}} = \frac{A_{\text{signed}} + 1}{2} \tag{6}$$

The reconstruction loss is:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} (\hat{A}_{ij} - A_{\text{target},ij})^2, \tag{7}$$

where $\mathcal{M} = \{(i,j) : A_{\text{target},ij} \neq 0.5\}$.

(2) Sign-Consistency Loss regulates the persona subspace to respect social alignment. Positively linked pairs should have close embeddings; negatively linked pairs should be separated by margin M:

$$\mathcal{L}_{\text{sign}} = \frac{1}{|E^+|} \sum_{(i,j)\in E^+} \|z_{p,i} - z_{p,j}\|_2^2 + \frac{1}{|E^-|} \sum_{(i,j)\in E^-} [\max(0, M - \|z_{p,i} - z_{p,j}\|_2)]^2 \tag{8}$$

(3) Belief Alignment Loss trains a classification head on the belief subspace. Given predicted logits $s_i = f_{\text{belief}}(z_{b,i})$ and labels $y_i \in \{0,1\}$:

$$\mathcal{L}_{\text{belief}} = -\frac{1}{N}\sum_{i=1}^{N} \big[y_i \log\sigma(s_i) + (1-y_i)\log(1-\sigma(s_i))\big] \tag{9}$$

(4) Orthogonality Loss enforces disentanglement by minimizing cross-covariance:

$$\mathcal{L}_{\text{orth}} = \left\| \frac{(z_b - \bar{z}_b)^\top (z_p - \bar{z}_p)}{N-1} \right\|_F^2, \tag{10}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\bar{z}_b, \bar{z}_p$ are mean-centered embeddings.

Model	Accuracy	Recall	Precision	F1-Score
roberta-large	**0.85**	**0.87**	**0.86**	**0.87**
google/gemini-2.0-flash	0.84	0.86	0.85	0.85
meta-llama/Llama-3.2-3B	0.83	0.84	0.83	0.83
roberta-base	0.80	0.83	0.81	0.82
deepset/gbert-base	0.83	0.81	0.82	0.82
microsoft/deberta-base	0.80	0.80	0.80	0.80
bert-large-uncased	0.77	0.77	0.77	0.77
distilbert-base-uncased	0.76	0.79	0.78	0.78
bert-base-uncased	0.75	0.77	0.76	0.77
aisingapore/SEA-LION-v1-7B	0.63	0.68	0.64	0.65
TelConGBERT(Pustet et al., 2024)	0.90	0.82	0.81	0.81
Goyal et al.(Goyal et al., 2025)	0.72	0.85	0.65	0.74
GPT-3.5	0.78	0.74	0.78	0.76
GPT-4	0.83	0.80	0.84	0.82
Llama 2	0.66	0.57	0.60	0.60

Table 2: Performance comparison of language models in detecting conspiratorial content, including additional fine-tuned, few-shot, and zero-shot models. Metrics reported are accuracy, recall, precision, and F1-score.

4.3 Stage 2(b): Narrative Archetype Discovery via Hierarchical Clustering

Having obtained disentangled persona embeddings $z_{p,i}{}_{i=1}^{N}$ from SiBeGNN, we identify distinct narrative archetypes through hierarchical clustering. These archetypes capture recurring communicative patterns defined by stylistic and behavioral traits rather than ideological content, allowing analysis of how different discourse modes coexist within the conspiratorial ecosystem.

Each message t_i is represented by its persona embedding $z_{p,i} \in \mathbb{R}^{d_p}$ learned in Stage 2(a), which encodes narrative style independent of belief polarity via the Sign-Disentanglement Loss. To reduce dimensionality while preserving semantic structure, Principal Component Analysis (PCA) is applied, retaining components that explain $p\%$ of the variance. This balances noise suppression with structural fidelity, enhancing clustering stability and interpretability. The reduced embeddings are clustered using agglomerative hierarchical clustering with Ward's linkage (Murtagh, 1983), which minimizes within-cluster variance while preserving hierarchical relationships.

The optimal number of clusters k^* is selected by evaluating candidate counts $k \in [k\_\min, k\_\max]$ using the silhouette coefficient (Shahapure and Nicholas, 2020), which measures cluster cohesion and separation. Post-hoc merging combines cluster pairs whose centroid cosine similarity exceeds a threshold τ, ensuring that final clusters correspond to semantically distinct narrative archetypes rather than minor stylistic variations.

To assess robustness and generalizability, we conducted ablation studies varying key hyperparameters: PCA variance retention ($pca\_var \in 0.5, 0.6, 0.7$), merge threshold ($merge\_th \in 0.75, 0.8$), and cluster bounds ($k\_\min \in 2, 3, 4,$

$k\_\max = 20$). Across 18 configurations, average coherence scores ranged from 0.360 to 0.386 (within 8% variation), demonstrating stable and interpretable clustering. All reported results use the optimal configuration from Table A2, confirming that the narrative archetype structure consistently emerges from the disentangled SiBeGNN embeddings rather than being an artifact of specific hyperparameter choices.

5 Results

Classification models: As shown in Table 2, RoBERTa-large achieved the highest performance (F1 = 0.866, accuracy = 0.852, precision = 0.86, recall = 0.87), while SEA-LION-v1-7B, a multilingual Southeast Asian model (AI Singapore, 2024), scored lowest (F1 = 0.653), suggesting that region-specific pretraining does not guarantee an advantage. RoBERTa-large was thus selected as the primary classifier and applied to the full corpus of 553,648 messages to generate binary conspiratorial labels.

Narrative Archetype Characterization and Clustering Quality: The seven narrative archetypes derived from SiBeGNN embeddings (Table 4) capture diverse conversational modes and user intents in Singapore-based Telegram groups. Larger clusters such as *General Discussions* and *Banking and Finance* emphasize everyday interaction, casual conversation, shopping, and financial literacy, reflecting socially grounded communication rather than overt ideological expression. *Contradictions in Authority* and *Medical Concerns* highlight institutional skepticism and health-related anxieties, while *General Legal Topics* reflects civic curiosity about rights and governance. Smaller clusters, *Group Moderation* and *Media Discussions*, center on meta-discourse regarding community norms and popular culture. Collectively, these archetypes illustrate that Telegram discourse spans civic inquiry, social bonding, and episodic skepticism, forming a continuum of everyday engagement rather than being dominated by conspiratorial narratives

LIWC (Boyd et al., 2022) analyses provide additional insight into cluster-specific linguistic characteristics. *General Legal Topics* employs practical, informational language with minimal emotion, reflecting public curiosity. *Medical Concerns* expresses personal experiences and anxieties with higher emotional and health-related language, including advocacy and empathy. *Media Discussions* exhibits informal, interactive language focused on entertainment and trending news. *Banking and Finance* uses analytical language centered on money, economics, risk, and skepticism, whereas *Contradictions in Authority* features elevated negative emotion and dissent, challenging institutional decisions through emotionally charged discourse. *Group Moderation* employs procedural, minimally emotional language, while *General Discussions* spans casual, inclusive exchanges on diverse everyday topics.

Conspiratorial discourse is distributed across these archetypes rather than confined to specific clusters. Elevated prevalence occurs in *Contradictions in Authority* and *Medical Concerns*, consistent with institutional skepticism and health policy critique. However, conspiratorial narratives also emerge in mundane clusters like *Banking and Finance* and *General Legal Topics*, where they occasionally intersect with systemic manipulation claims. This distribution reinforces the central finding that conspiratorial discourse operates within everyday communicative practices, emphasizing the need for context-aware platform governance rather than blanket removal policies.

Clustering quality was assessed using topic coherence, silhouette score, and the Davies–Bouldin index, integrated into a composite cDBI (Krasnov and Sen, 2019) as shown in Table 3:

$$\text{cDBI} = \frac{\text{Davies–Bouldin Index}}{\text{Average Coherence}}. \qquad (11)$$

Lower cDBI indicates better joint coherence–compactness performance. Hierarchical clustering on SiBeGNN embeddings achieved the lowest cDBI (8.38), outperforming Bertopic (13.60) and other methods (33.13–67.27), demonstrating that the latent-graph hierarchical approach produces semantically coherent, well-separated clusters. Ablation studies removing L_{orth}, L_{sign}, or L_{belief} increased cDBI by 58.2%, 94.6%, and 136.9%, respectively, while using only L_{recon} caused a 271.4% increase. These results confirm that all four loss components are essential for generating interpretable, semantically meaningful narrative archetypes, reflecting both content and structural topology of the discourse.

Manual verification of narrative archetypes: To evaluate the ability of different clustering algorithms to capture the diversity of narrative archetypes, we manually compared the clusters

Model	Avg. Coherence	Silhouette Score	Davies–Bouldin Index	cDBI
Hierarchical clustering with SiBeGNN embeddings	0.386	-0.021	3.233	**8.38**
Bertopic clusters with vanilla Roberta-large embeddings	0.331	-0.014	4.502	13.60
Bertopic clustering with SiBeGNN embeddings	0.196	-0.042	6.494	33.13
Hierarchical clusters with vanilla Roberta-large embeddings	0.234	-0.015	8.246	35.24
HBScan clusters with vanilla Roberta-large embeddings	0.235	-0.018	13.360	56.85
Gaussian mixture model with SiBeGNN embeddings	0.207	-0.015	13.924	67.27
Spectral clustering with vanilla Roberta-large embeddings	0.218	-0.019	12.345	56.63
KMeans with SiBeGNN embeddings	0.189	-0.025	14.567	72.10

Table 3: Comparison of clustering quality metrics using cDBI. Lower cDBI values indicate better joint coherence–compactness performance.

Narrative Archetype	Description	Common Keywords	# Messages	Avg. Length	# Conspiratorial
Legal Topics	Everyday legal issues, rights, exemptions, and practical legality.	commoner, legalized, crime, deemed, run, definitely	69,357	160.22	3,468
Medical Concerns	Personal feelings on medical topics and advocacy for individual rights.	place, medical, feel, people, concerned, individuals	11,357	170.36	3,407
Media Discussions	Commentary on pop culture, concerts, news stories, and viral events.	taylor, cruel, concert, attend, swift, 81	41,625	293.33	2,081
Banking and Finance	Analysis of finance, banking, rates, and the economic system.	rates, cdp, account, fed, likely, till	70,726	232.20	14,145
Contradictions in Authority / Riot	Opinions on authority, health policy, and contradictions from officials.	contradictory, gotta, dun, boss, totally, spread	55,197	137.33	38,638
Group Moderation	Moderation, rules, respectful conduct, and group management.	disrespectful, expletives, explicitly, irrelevant, disappears, exists	5,416	36.41	271
General Discussions	Broad conversations on casual chat, shopping, food, and miscellaneous events.	shopping, areas, jb, talk, abt, non, just, like, yes	299,470	125.47	44,921

Table 4: Narrative archetype clusters with estimated conspiratorial message counts. Total estimated conspiratorial messages: 106,931.

Model	Ground Truth Overlap (/9)
Hierarchical clustering with SiBeGNN embeddings	5/9
Bertopic clusters with vanilla RoBERTa-large embeddings	7/9
Bertopic clustering with SiBeGNN embeddings	6/9
Hierarchical clusters with vanilla RoBERTa-large embeddings	4/9
HBScan clusters with vanilla RoBERTa-large embeddings	2/9
Gaussian mixture model with SiBeGNN embeddings	5/9
Spectral clustering with vanilla RoBERTa-large embeddings	5/9
KMeans with SiBeGNN embeddings	5/9

Table 5: Comparison of narrative archetype coverage across clustering methods. Scores indicate the number of archetypes captured out of 9 possible ground truth narrative archetypes

produced by each method against nine ground-truth archetypes defined in our study (Table 5). Archetypes were defined by two experts who had published at least five articles on misinformation in Singapore and the broader region. Experts developed archetypes independently, and a third expert resolved disagreements. These archetypes represented distinct core narratives in the news and online environment during the study period. For instance, *A1: Technocratic Control vs Personal Autonomy* reflects governance systems overriding individual choice, signaled by mentions of mandates, VDS compliance, surveillance, and bodily control. *A2: Betrayal by Trusted Institutions* denotes perceived deception or incompetence by protective institutions, exemplified by government distrust or hospitals hiding data. *A3: Unequal Citizenship & Social Stratification* captures societal tiers with unequal rights, such as vaxxed versus unvaxxed populations and exclusion from jobs or venues. *A4: Elite Collusion & Loss of National Sovereignty* highlights decision-making dominated by elites or external powers, including global institutions or corporate capture. *A5: Information Manipulation & Manufactured Consent* focuses on media shaping opinion through censorship or framing, for example, news media distrust or propaganda. *A6: Moral / Civilisational Decline* reflects the erosion of ethical boundaries, often involving religious framing, children, consent, or dehumanization. *A7: Awakening, Resistance & Minority Identity* describes minorities claiming awareness of hidden truths, signaled by terms like "sheeple," "awakened," or "truth-seekers." *A8: Everyday Life Under Systemic Stress* emphasizes the impact of large systems on daily life, such as disruptions to food, jobs, travel, healthcare access, or cost of living. Finally, *A9: External Conflict as Moral Contrast* captures the use of foreign conflicts to interpret local events, for

example, comparisons involving the US, China, or Russia. The manual verification results show that **Hierarchical clustering with SiBeGNN embeddings** captured five of the nine archetypes, successfully identifying both everyday and conspiratorial discourse clusters and demonstrating strong separation of narrative personas. **Bertopic clusters with vanilla RoBERTa-large embeddings** captured seven archetypes, showing higher semantic coverage but lower interpretability due to some cluster overlap. **Bertopic clustering with SiBeGNN embeddings** identified six archetypes, balancing persona representation with semantic alignment. **Hierarchical clusters with vanilla RoBERTa-large embeddings** captured four archetypes, missing several conspiratorial narratives such as A2 and A6. **HBScan clusters with vanilla RoBERTa-large embeddings** captured only two archetypes, indicating poor granularity and failure to differentiate stylistic nuances. Other methods, including **Gaussian mixture model with SiBeGNN embeddings**, **Spectral clustering with vanilla RoBERTa-large embeddings**, and **KMeans with SiBeGNN embeddings**, captured five archetypes each, demonstrating moderate coverage but lower interpretability relative to hierarchical approaches. These results indicate that **SiBeGNN-based hierarchical clustering** provides the best balance between coverage and interpretability. By leveraging disentangled, sign-aware embeddings, this approach effectively distinguishes both conspiratorial and everyday discourse modes, producing coherent and semantically meaningful narrative archetypes. In contrast, clustering methods that rely solely on vanilla embeddings or standard algorithms either fail to cover key archetypes or produce overlapping clusters, underscoring the importance of using belief-aware, persona-disentangled representations to capture the full spectrum of online discourse narratives.

6 Discussion and Implications

This study shows that conspiratorial discourse in Singapore-based Telegram groups is embedded within everyday communication rather than isolated spaces, challenging prevailing assumptions about online radicalization and highlighting the need to study such narratives within broader social discourse.

Methodological implications: Our signed belief graph–based hierarchical clustering achieves the lowest cDBI (8.38) compared to standard methods (13.60–67.27), indicating more interpretable narrative archetypes (Krasnov and Sen, 2019). The Sign-Disentanglement Loss effectively separates belief polarity from stylistic variation, extending beyond conspiracy detection to applications in stance detection, political discourse analysis, and other tasks requiring joint modeling of relational polarity and semantic content.

Platform governance and generalizability: The seven archetypes exhibit distinct conspiracy prevalence patterns, indicating that moderation should be archetype-aware. For example, *Medical Concerns* may benefit from targeted health interventions, while *Contradictions in Authority* may require approaches addressing institutional trust deficits. Although the framework is platform- and region-agnostic, the identified archetypes reflect Singapore's socio-political context, underscoring the need for cross-platform and cross-regional studies to distinguish universal from context-specific patterns.

7 Conclusion

This study shows that conspiratorial discourse in Singapore-based Telegram groups exists within everyday communication, spanning legal, health, media, and financial discussions, rather than in isolated echo chambers. Using a two-stage framework combining transformer-based classification and Signed Belief Graph Neural Networks (SiBeGNN), we identified seven narrative archetypes, revealing how conspiratorial content interweaves with ordinary social interaction. The findings challenge assumptions about online radicalization: narrative archetypes from *General Legal Topics* to *Contradictions in Authority* illustrate that pragmatic, affective, and ideological communication coexist. Even skepticism-laden clusters appear within broader conversational contexts, highlighting the need for interventions that consider conspiratorial discourse's embeddedness in everyday digital life. Methodologically, the Sign-Disentanglement Loss separates belief polarity from narrative style, yielding superior clustering quality (cDBI = 8.38). This approach offers a replicable framework for studying belief-driven discourse across platforms, integrating signed network analysis with transformer embeddings to capture semantic content and relational structure, with applications in stance detection, and political discourse analysis.

8 Limitations

This study has several limitations presenting opportunities for future research. First, our analysis relies on textual embeddings from a specific transformer model; alternative architectures or multimodal data may reveal different patterns. Second, our normalization and weighting schemes, while theoretically justified, remain heuristic and could be optimized for specific objectives. A critical limitation is our reliance on text-based features for archetype discovery. Narrative archetypes emerge from richer behavioral, temporal, and network-level dynamics beyond text alone. Notably, *lurkers*, users who consume content without contributing, remain invisible to text-based analyses yet constitute substantial portions of communities and indirectly influence information ecosystems. Capturing such latent participation requires incorporating posting frequency, temporal activity patterns, engagement metrics, response latency, and structural network position (e.g., central versus peripheral actors). Future extensions incorporating multimodal or interaction-based signals could better model latent behavioral dimensions underlying archetype formation. Finally, our dataset is restricted to Singapore-based Telegram groups and may not generalize to other regions or platforms with different cultural contexts, regulatory environments, or community norms. Future work could incorporate temporal dynamics of cluster evolution (Shimgekar et al., 2025b,a), integrate engagement metrics and social network structure, and conduct cross-platform comparative studies to identify universal versus context-specific patterns.

9 Ethical considerations

We implemented privacy safeguards including paraphrasing, anonymization, and analysis at the aggregate level. We acknowledge risks of surveillance misuse and emphasize that institutional skepticism is a core element of democratic discourse, distinct from false misinformation. Labeling health concerns as "conspiratorial" risks pathologizing legitimate grievances; our annotation protocols distinguish evidence-free claims from evidence-based criticism, though this boundary is context-dependent. We commit to public release of methods and code while restricting access to raw message data. The dataset may contain offensive or harmful language reflective of real-world discourse; we do not amplify such content and restrict access to the data in accordance with institutional ethical guidelines.

The labelling task required labeling each message as conspiratorial or non-conspiratorial based solely on message content, without attempting to infer author intent or user identity. As annotators were domain experts and collaborators on the project, no formal consent process or risk disclaimer was required; however, annotators were informed that the data may contain offensive or sensitive language typical of online political discourse.

References

Rachel Abdou and 1 others. 2021. Good governance and social-media overload in singapore. *Journal of Asian Public Policy*. Related study on digital governance and information overload.

AI Singapore. 2024. SEA-LION-v1-7B: A Southeast Asian Multilingual Language Model. https://huggingface.co/aisingapore/SEA-LION-v1-7B. Accessed: 2025-10-07.

Val Alvern Cueco Ligo, Lam Yin Cheung, Roy Ka-Wei Lee, Koustuv Saha, Edson C Tandoc Jr, and Navin Kumar. 2025. User archetypes and information dynamics on telegram: Covid-19 and climate change discourse in singapore. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2685–2688.

Mathias Angermaier, João Pinheiro-Neto, Elisabeth Hoeldrich, and Jana Lasser. 2025. The schwurbelarchiv: a german language telegram dataset for the study of conspiracy theories. arXiv preprint arXiv:2504.06318.

Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2019. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. volume 13, pages 496–507. IEEE.

Md Rabiul Awal, Minh Dang Nguyen, Roy Ka-Wei Lee, and Kenny Tsu Wei Choo. 2022. Muscat: Multilingual rumor detection in social media conversations. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 455–464. IEEE.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Keyu Chen, Marzieh Babaeianjelodar, Yiwen Shi, Kamila Janmohamed, Rupak Sarkar, Ingmar Weber, Thomas Davidson, Munmun De Choudhury, Jonathan Huang, Shweta Yadav, and 1 others. 2023a. Partisan us news media representations of syrian refugees. In *Proceedings of the International AAAI*

Conference on Web and Social Media, volume 17, pages 103–113.

Qiang Chen, Wei Zhang, and Huan Li. 2023b. Categorizing misinformation and conspiratorial content during covid-19 on social media. *Computers in Human Behavior Reports*, 3:100121.

Daniel Wai Kit Chin, Kwan Hui Lim, and Roy Ka-Wei Lee. 2024. Rumorgraphxplainer: Do structures really matter in rumor detection. *IEEE Transactions on Computational Social Systems*, 11(5):6038–6055.

Avihay Chriqui and Inbal Yahav. 2022. Hebert and hebemo: A hebrew bert model and a tool for polarity analysis and emotion recognition. *INFORMS Journal on Data Science*, 1(1):81–95.

Tyler Derr, Yao Ma, and Jiliang Tang. 2018. Signed graph convolutional network. *arXiv preprint arXiv:1808.06354*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

Ahmad Diab, Rr Nefriana, and Yu-Ru Lin. 2024. Classifying conspiratorial narratives at scale: False alarms and erroneous connections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 340–353.

Karen M Douglas, Robbie M Sutton, and Aleksandra Cichocka. 2017. The psychology of conspiracy theories. *Current directions in psychological science*, 26(6):538–542.

Andy Egan and Brian Weatherson. 2011. Epistemic modality.

Susmita Gangopadhyay, Danilo Dessi, Dimitar Dimitrov, and Stefan Dietze. 2025. Telescope: A longitudinal dataset for investigating online discourse and information interaction on telegram. arXiv preprint arXiv:2504.19536.

Google DeepMind. 2024. Gemini 2.0 flash: Scaling multimodal foundation models for efficiency and responsiveness. https://deepmind.google/technologies/gemini/. Accessed: 2025-10-16.

Abhay Goyal, Val Alvern Cueco Ligo, Lam Yin Cheung, Roy Ka-Wei Lee, Koustuv Saha, Edson C Tandoc Jr, and Navin Kumar. 2025. Analyzing conspiratorial content across singapore-based telegram groups. *medRxiv*, pages 2025–07.

Abhay Goyal, Muhammad Siddique, Nimay Parekh, Zach Schwitzky, Clara Broekaert, Connor Michelotti, Allie Wong, Lam Yin Cheung, Robin O Hanlon, Munmun De Choudhury, and 1 others. 2023. Chatgpt and bard responses to polarizing questions. *arXiv preprint arXiv:2307.12402*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

X. Hu, Jiliang Tang, Yao Ma, and Yu Wang. 2024. Disentangled generative graph representation learning (diggr). arXiv preprint arXiv:2408.13471.

Fedor Krasnov and Anastasiia Sen. 2019. The number of topics optimization: Clustering approach. *Machine Learning and Knowledge Extraction*, 1(1):25.

Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. 2016. Edge weight prediction in weighted signed networks. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 221–230. IEEE.

H. Li, H. Peng, Z. Liu, L. He, J. Wu, and S. Yu Philip. 2021. Disentangled graph contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jian Ma, Jiliang Tang, Qiaozhi Zhu, and Qiaozhu Mei. 2019a. Disentangled graph convolutional networks (disengcn). In *Proceedings of Machine Learning Research, Vol. 97*, pages 5–19.

Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019b. Disentangled graph convolutional networks. In *International conference on machine learning*, pages 4212–4221. PMLR.

Raija Markkanen and Hartmut Schröder. 1997. Hedging: A challenge for pragmatics and discourse analysis.

Meta AI. 2024. The llama 3.2 model family. https://ai.meta.com/llama/. Accessed: 2025-10-16.

Fionn Murtagh. 1983. *A Survey of Recent Advances in Hierarchical Clustering Algorithms*, volume 26. Oxford University Press.

Yashna Nainani, Kaveh Khoshnood, Ashley Feng, Muhammad Siddique, Clara Broekaert, Allie Wong, Koustuv Saha, Roy Ka-Wei Lee, Zachary M Schwitzky, Lam Yin Cheung, and 1 others. 2022. Categorizing memes about abortion. In *The International Conference on Weblogs and Social Media*.

Wee Ng and Chong Loke. 2020. Analyzing health-policy and personal experience discourse on telegram in singapore. *BMC Public Health*.

Milena Pustet, Elisabeth Steffen, and Helena Mihaljević. 2024. Detection of conspiracy theories beyond keyword bias in german-language telegram using large language models. *arXiv preprint arXiv:2404.17985*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.

Soorya Ram Shimgekar, Abhay Goyal, Shayan Vassef, Koustuv Saha, Christian Poellabauer, Xavier Vautier, Pi Zonooz, and Navin Kumar. 2025a. Nimblelabs: Accelerating healthcare ai development through agentic ai. *preprints.org preprints:202508.1713.v1*.

Soorya Ram Shimgekar, Shayan Vassef, Abhay Goyal, Navin Kumar, and Koustuv Saha. 2025b. Agentic ai framework for end-to-end medical data inference. *arXiv preprint arXiv:2507.18115*.

Aelita Skarzauskiene, Monika Maciuliene, Aiste Dirzyte, and Gintare Guleviciute. 2025. Profiling antivaccination channels in telegram: early efforts in detecting misinformation. *Frontiers in Communication*, 10:1525899.

Edson C. Jr. Tandoc, Kaidi Jenkins, and Rachel Neo. 2021. Developing a perceived social media literacy scale: Evidence from singapore. *International Journal of Communication*, 15:16118.

Aleksandra Urman and Stefan Katz. 2020. What they do in the shadows: examining the far-right networks on telegram. *Information, Communication & Society*, 25(7):904–923.

Joseph E Uscinski and Joseph M Parent. 2014. *American conspiracy theories*. Oxford University Press.

Sallie van der Linden, Van Hien, and Rachel Neo. 2023. Fatherhood, social media penetration, and information regulation in singapore. *Asian Journal of Communication*. In press.

Samantha Walther and Andrew McCoy. 2021. Us extremism on telegram: Fueling disinformation, conspiracy theories, and accelerationism. *Perspectives on Terrorism*, 15(2):41–60.

Jing Zeng, Mike S Schäfer, and Thaiane M Oliveira. 2022. Conspiracy theories in digital environments: Moving the research field forward. *Convergence*, 28(4):929–939.

Ziwei Zhang, Chuan Shi, Jieyu Yang, and Yao Ma. 2024. Signed graph representation learning: A survey. *arXiv preprint arXiv:2402.15980*.

Ziwei Zhang, Jieyu Yang, and Chuan Shi. 2021. Signed graph neural network with latent groups (gs-gnn). In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*.

Fabiana Zollo, Alessandro Bessi, Michela Del Vicario, Antonio Scala, Guido Caldarelli, Louis Shekhtman, Shlomo Havlin, and Walter Quattrociocchi. 2017. Debunking in a world of tribes. *PloS one*, 12(7):e0181821.

A Appendix

Text (masked for brevity)	Label
Zionism – A political movement that sucks the world dry. The Synagogue of Satan consists of Jews and non-Jews. (Biblical terminology hijacked strategically by Kazarian mafia/Mossad)	1
Long list of hawker centres temporarily closed for cleaning. Some users speculate the word “affected” may imply closures targeting unvaccinated patrons; others verify on NEA website that closures are for cleaning only. Debate arises over tone and intention. [...]	0

Table A1: Masked chat excerpts with binary conspiratorial labels. Each example is labeled as conspiratorial (1) or not (0), with unimportant text masked using [...].

PCA	Th.	k_{min}	k_{max}	Batch	State	Coher.
0.5	0.75	2	20	64	42	0.372
0.5	0.75	3	20	64	42	0.381
0.5	0.75	4	20	64	42	0.365
0.5	0.80	2	20	64	42	0.388
0.5	0.80	3	20	64	42	0.379
0.5	0.80	4	20	64	42	0.384
0.6	0.75	2	20	64	42	0.376
0.6	0.75	3	20	64	42	0.363
0.6	0.75	4	20	64	42	0.382
0.6	**0.80**	**2**	**20**	**64**	**42**	**0.386**
0.6	0.80	3	20	64	42	0.374
0.6	0.80	4	20	64	42	0.386
0.7	0.75	2	20	64	42	0.368
0.7	0.75	3	20	64	42	0.383
0.7	0.75	4	20	64	42	0.360
0.7	0.80	2	20	64	42	0.371
0.7	0.80	3	20	64	42	0.378
0.7	0.80	4	20	64	42	0.372

Table A2: Ablation results for hierarchical clustering parameters. Coherence values remain within 0.360–0.390, indicating stable behavior across parameter settings.

Loss Variant	Avg. Coherence	Silhouette Score	Davies–Bouldin Index	cDBI	Δ cDBI
Full SiBeGNN (all losses)	0.386	-0.021	3.233	**8.38**	—
w/o $\mathcal{L}_{\text{orth}}$ (no orthogonality)	0.341	-0.028	4.521	13.26	+58.2%
w/o $\mathcal{L}_{\text{sign}}$ (no sign consistency)	0.318	-0.032	5.187	16.31	+94.6%
w/o $\mathcal{L}_{\text{belief}}$ (no belief alignment)	0.297	-0.037	5.894	19.85	+136.9%
w/o $\mathcal{L}_{\text{orth}}$ & $\mathcal{L}_{\text{sign}}$ (both removed)	0.289	-0.041	6.523	22.57	+169.3%
Only $\mathcal{L}_{\text{recon}}$ (no task supervision)	0.251	-0.046	7.812	31.12	+271.4%
Vanilla RoBERTa (no GNN)	0.234	-0.015	8.246	35.24	+320.5%

Table A3: Ablation study of sign-disentanglement loss components. Each row reports performance when specific loss terms are removed. Δ cDBI denotes percentage degradation relative to the full model. Lower cDBI values indicate better clustering quality.