Belief Is All You Need: Signed Belief Graph Neural Networks for Topic Modeling in Conspiratorial Discourse

Soorya Ram Shimgekar[1,2] Abhay Goyal[2] Roy Ka-Wei Lee[3] Koustuv Saha[1]
Pi Zonooz[2] Edson C Tandon Jr[4] Navin Kumar[2]

[1]University of Illinois Urbana-Champaign, USA
[2]Nimblemind, USA
[3]Singapore University of Technology and Design, Singapore
[4]Nanyang Technological University, USA

Abstract

Conspiratorial discourse is increasingly present in online communication, yet how it is organized across discussion topics remains unclear. We analyze Singapore-based Telegram groups to examine how conspiratorial content appears within everyday conversations rather than isolated echo chambers. To better capture the structure of such discussions, we propose a two-stage framework for topic modeling tailored to conspiratorial posts. First, a RoBERTa-large classifier identifies conspiratorial messages (F1 = 0.866) using 2,000 expert-annotated examples. We then construct a graph where connections reflect textual similarity and conspiratorial stance. This graph is modeled using a Signed Belief Graph Neural Network (SiBeGNN), which learns message embeddings that distinguish conspiratorial from non-conspiratorial content. We apply hierarchical clustering on these embeddings to perform topic modeling over 553,648 Telegram messages, producing seven topic clusters: General Legal Topics, Medical Concerns, Media Discussions, Banking and Finance, Contradictions in Authority, Group Moderation, and General Discussions. Our method substantially outperforms standard embedding-based clustering approaches (cDBI = 8.38 vs. 13.60–67.27), with manual evaluation showing 88% inter-rater agreement in cluster interpretation. The results show that conspiratorial content appears across multiple everyday topics, including finance, law, and daily life, rather than forming isolated thematic communities.

1 Introduction

The Web has transformed how information is produced and circulated, enabling narratives to spread rapidly across digital communities while creating new challenges for information integrity and public trust. Among the most persistent forms of online narratives are conspiratorial explanations, which attribute major social, political, or health events to covert actors pursuing hidden and malevolent agendas (Zeng et al., 2022). These explanations often challenge official accounts and frame institutions, governments, or corporations as deliberately deceptive. Prior research shows that conspiratorial discourse can undermine trust in science and governance, intensify political polarization, and accelerate the spread of misinformation during periods of crisis (Douglas et al., 2017; Uscinski and Parent, 2014; Zollo et al., 2017). Understanding how conspiratorial content is organized and discussed within online communities is therefore an important problem for web science and computational social science.

A central challenge in studying conspiratorial discourse is identifying the topics and discussion themes through which such narratives appear in everyday online communication. Traditional topic modeling and clustering methods rely primarily on semantic similarity between texts (Cheng et al., 2023; Xiao et al., 2026). However, conspiratorial messages often share vocabulary/semantics with non-conspiratorial discussions, particularly in domains such as health, politics, or economics (Hoseini et al., 2021; Reiter-Haas et al., 2024). As a result, standard semantic clustering approaches frequently group conspiratorial and non-conspiratorial messages together, making it difficult to isolate how conspiratorial discourse is structured across topics. This limitation motivates the need for topic modeling approaches that account not only for textual similarity but also for differences in conspiratorial stance or belief expression.

Digital platforms shape how conspiratorial discourse emerges and spreads. Telegram has become

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 341–355

a prominent venue due to its hybrid structure of private messaging, large public groups, and relatively minimal moderation, enabling rapid narrative diffusion and loosely connected communities sharing counter-mainstream perspectives (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 developments, often expressing skepticism toward institutional authority (Ng and Loke, 2020). Singapore's media environment combines high digital connectivity with strong information governance, where regulated discourse coexists with relatively unmoderated messaging platforms (Van et al., 2023; Tandoc et al., 2021). Large-scale efforts such as *Schwurbelarchiv* (Angermaier et al., 2025) and *TeleScope* (Gangopadhyay et al., 2025) highlight the value of analyzing Telegram ecosystems but also the need for methods that identify coherent discussion topics within conspiratorial discourse.

We address this gap by proposing a belief-aware topic modeling framework that combines transformer-based classification, signed graph representation learning, and hierarchical clustering to identify discussion topics in conspiratorial Telegram conversations. In this study, we analyze Singapore-based Telegram groups to examine how conspiratorial discourse appears within a low-moderation environment. Beyond identifying conspiratorial messages, we study how they are embedded within broader conversational topics and interactions in public group discussions. Using embeddings learned from a Signed Belief Graph Neural Network (SiBeGNN), we perform hierarchical clustering to organize messages into coherent topics. This approach allows us to capture discussion themes that better reflect conspiratorial discourse while preserving the broader conversational context in which these messages appear.

We ask: How is conspiratorial discourse organized across discussion topics in Singapore-based Telegram groups, and what themes emerge from message-level patterns?

This work makes three main contributions. **First**, we propose a belief-aware topic modeling framework (SiBeGNN) for analyzing conspiratorial discourse in large-scale Telegram conversations.

Second, we introduce a signed belief graph where edges encode both textual similarity and agreement or disagreement in conspiratorial labels. SiBeGNN learns message embeddings that separate conspiratorial stance from general discourse characteristics, improving representations for topic discovery.

Third, we apply the framework to 553,648 messages from Singapore-based Telegram groups and identify seven discussion topics spanning domains such as legal issues, health concerns, media discussions, and financial conversations.

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 covert and malevolent groups, often emerging from 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, 2021; Tandoc et al., 2021). While this literature explains why conspiratorial beliefs emerge and persist, it provides limited insight into how conspiratorial discourse is organized across topics within online platforms. In particular, existing studies often overlook how platform-level affordances and discussion structures shape the thematic organization of conspiratorial conversations (Chen et al., 2023; Goyal et al., 2023).

Telegram as a Platform for Conspiratorial Discourse: Telegram has become a major venue for conspiratorial and extremist communication due to its hybrid public–private structure, large group capacities, and relatively minimal moderation (Urman and Katz, 2020; Skarzauskiene et al., 2025). Features such as message forwarding, channel interconnectivity, and broadcast groups enable rapid narrative diffusion and loosely connected communities sharing counter-mainstream perspectives (Nainani et al., 2022; Chen et al., 2022). Large-scale datasets such as *Schwurbelarchiv* (Angermaier et al., 2025) and *TeleScope* (Gangopadhyay et al., 2025) have enabled multimodal and longitudinal analysis of Telegram conspiratorial ecosystems. However, most work focuses on Euro-American contexts, leaving regional linguistic and cultural dynamics underexplored (Ng and Loke, 2020; Goyal et al., 2025). We address this gap by analyzing conspiratorial discourse in Singapore-based Telegram groups.

Computational Detection of Conspiratorial Content: Transformer-based models such as BERT,

RoBERTa, and DeBERTa, along with newer systems like LLaMA and SEA-LION, have advanced automated detection of conspiratorial content (Urman and Katz, 2020; Skarzauskiene et al., 2025). Despite strong benchmark performance, these models often degrade under domain shift, multilingual noise, and informal online language (Walther and McCoy, 2021; Alvern Cueco Ligo et al., 2025). Moreover, most work treats detection as binary classification, focusing on identifying conspiratorial messages rather than understanding how conspiratorial discussions organize across broader topics.

Graph Neural Networks for Social Discourse: Graph neural networks naturally model relational structures in online discourse. Signed graph neural networks represent positive and negative relationships between nodes, capturing agreement and disagreement consistent with structural balance theory (Derr et al., 2018; Zhang et al., 2021, 2024). Prior work typically focuses on explicit social links or sentiment signals and rarely models belief-oriented structures in large-scale text networks. Most approaches also fail to separate factors shaping message representations, such as belief stance and stylistic variation. Our approach addresses this limitation by modeling message relationships using a Signed Belief Graph Neural Network.

Disentangled Representation Learning: Disentangled representation learning separates distinct generative factors in embeddings. Graph-based approaches such as DisenGCN (Ma et al., 2019a) and DiGGR (Hu et al., 2024) show how orthogonal embedding spaces can capture multiple structural patterns in graph data. However, these methods have not been adapted to signed graphs representing belief relationships between textual messages. Integrating disentangled learning with signed graph representations enables embeddings that capture both semantic similarity and belief differences.

Clustering and Topic Modeling of Conspiratorial Discourse: Clustering and topic modeling uncover latent thematic structure in large text corpora (Li et al., 2021). Methods such as hierarchical clustering, HDBScan, Gaussian mixture models, and neural topic models group semantically similar messages into discussion topics. However, they rely primarily on textual similarity and typically ignore belief polarity or antagonistic relationships between messages. This limitation is critical for conspiratorial discourse, where messages may span diverse subjects but share skepticism toward institutions (Uscinski and Parent, 2014). We address this by combining belief-aware embeddings with hierarchical clustering to identify discussion topics within conspiratorial Telegram conversations.

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 topic modeling and clustering analysis (Stage 2).

Stage 1: Annotated Training Set. We randomly sampled 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. The messages were extracted using Telegram API. A message was labeled as conspiratorial only if it hypothesized the existence of a hidden, coordinated group pursuing a malevolent goal, rather than merely expressing disagreement with official accounts or uncertainty. In making these determinations, annotators explicitly considered the presence of structural elements such as actors, goals, and strategies. Ambiguous cases were resolved through calibration and consensus. A detailed comparison of conspiratorial and non-conspiratorial criteria is provided in Table A2.

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 construction of a signed belief graph where edges encode textual similarity and conspiratorial stance alignment. SiBeGNN then learns message embeddings from this graph, which are used for hierarchical clustering to identify discussion topics in the corpus.

4 Methods

We adopt a two-stage framework. **First**, we fine-tune RoBERTa-large to classify messages as conspiratorial or non-conspiratorial. **Second**, we con-

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.

struct a signed belief graph in which nodes represent messages, edge signs encode conspiratorial agreement or disagreement, and edge weights capture textual similarity. This graph is modeled using our Signed Belief Graph Neural Network (SiBeGNN), which learns message embeddings that capture both semantic similarity and differences in conspiratorial stance. The learned embeddings are then used for hierarchical clustering to perform topic modeling over the corpus, producing seven discussion topics. An overview of the full pipeline, along with illustrative examples, is shown in Figure A1.

4.1 Stage 1: Conspiratorial Content Classification

The first stage of our framework trains a binary classifier to distinguish conspiratorial from non-conspiratorial messages. To select the most suitable model, we compared nine transformer-based architectures commonly used for discourse classification: 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). All models were fine-tuned on the same balanced dataset of 2,000 manually annotated messages using identical preprocessing and tokenization pipelines to ensure fair comparison.

Training was performed for 3 epochs with a learning rate of 2×10^{-5} and batch size of 16, with evaluation after each epoch. The annotated dataset was split into training and test sets using an 80/20 stratified split with $k = 5$. Model performance was evaluated using accuracy, precision, recall, and F1-score on the held-out test set. The best-performing classifier was then applied to the full corpus of 553,648 messages to assign binary labels (conspiratorial vs. non-conspiratorial). Example messages are shown in Table A1, and hyperparameter ablation results are reported in Table A3.

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 are first cleaned and encoded to obtain semantic representations. For each message, we construct an *information vector* that combines semantic content with features related to belief expression. Semantic content(s_i) is represented using `sentence-transformers/all-MiniLM-L6-v2` (Reimers and Gurevych, 2019), producing a dense 384-dimensional embedding.

To capture belief expression, we extract four discourse-level features inspired by prior work on belief and stance (van Dijk, 1998; Somasundaran and Wiebe, 2010). **Epistemic modality** (e_i) measures certainty versus hedging in language (Markkanen and Schröder, 1997). **Agency** (a_i) captures whether responsibility is expressed through active or passive framing (Somasundaran and Wiebe, 2010). **Sentiment polarity** (p_i) represents evaluative stance using a transformer-based sentiment model mapped to $[-1, 1]$ (Chriqui and Yahav, 2022; Choi et al., 2016). **Emotion spectra** (m_i) encode probabilities over core emotions (joy, anger, fear, sadness) using a pretrained emotion model (Araque et al., 2019, 2017).

$$x_i = \text{Norm}\left([\, s_i \,;\, e_i \,;\, a_i \,;\, p_i \,;\, m_i \,]\right) \quad (1)$$

where $[\cdot;\cdot]$ denotes vector concatenation and $\text{Norm}(\cdot)$ represents mean-centering followed by L2 normalization. A detailed explanation of these features is provided in Table A4.

Using the message information vectors x_i, we construct a signed graph $G = (V, E, W, S)$, where each node corresponds to a message. Edge weights $W_{ij} \in [0, 1]$ represent cosine similarity between message information vectors x_i and x_j, while edge signs $S_{ij} \in \{+1, -1\}$ encode belief alignment: positive edges connect messages with the same conspiratorial label, and negative edges connect messages with opposing labels. To maintain sparsity, edges are retained only when W_{ij} exceeds 0.5.

The resulting signed belief graph captures both semantic similarity and belief polarity in discourse, enabling the model to represent agreement and opposition between messages and to identify topics that reinforce or challenge conspiratorial narratives (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 message representations from the signed graph. In conspiratorial discussions, messages may discuss the same topic while expressing different viewpoints (e.g., supporting or rejecting conspiratorial claims). A single embedding often mixes these signals, making it difficult to separate conspiratorial stance from general discourse patterns.

To address this, SiBeGNN learns two embedding components for each node v_i, initialized with its message information vector x_i. The first component $z_{b,i} \in \mathbb{R}^{d_b}$ captures whether a message aligns with conspiratorial or non-conspiratorial viewpoints. The second component $z_{d,i} \in \mathbb{R}^{d_d}$ captures general discourse characteristics of the message, such as writing style independent of conspiratorial stance. The final message representation is obtained by concatenating the two components, $z_i = [z_{b,i}; z_{d,i}] \in \mathbb{R}^{d_b+d_d}$.

Architecture: Each node is initialized with its message information vector, forming the matrix $X \in \mathbb{R}^{N\times d_{\text{in}}}$. The model applies two graph convolution layers to propagate information through the signed graph: one over positive edges E^+ and one over negative edges E^-. Message passing is defined as:

$$\begin{aligned} h^+ &= \text{ReLU}(\text{GCNConv}^+(X, E^+)), \\ h^- &= \text{ReLU}(\text{GCNConv}^-(X, E^-)), \end{aligned} \tag{2}$$

where GCNConv denotes a graph convolution operation. The resulting representations are concatenated and projected to produce a shared hidden representation:

$$h = \text{ReLU}(W_h[h^+; h^-]), \tag{3}$$

where W_h is a learnable weight matrix. From this representation, the model produces two outputs through separate linear projections:

$$z_b = W_b h, \quad z_d = W_d h, \tag{4}$$

where z_b captures the conspiratorial stance of a message and z_d captures general discourse characteristics independent of conspiratorial stance. Both embeddings are L2-normalized for stability. This design allows the model to propagate information differently across agreement and disagreement edges while learning message representations.

Sign-Disentanglement Loss: SiBeGNN is trained using the Adam optimizer with weight decay, and the checkpoint with the lowest total loss is selected. After training, the model produces two embeddings for each message: $z_{b,i}$ capturing conspiratorial stance, and $z_{d,i}$ capturing discourse characteristics independent of stance. These embeddings are used for hierarchical clustering in Stage 2(b). Training uses a composite loss with four objectives to preserve graph structure, maintain stance consistency, and encourage the two embeddings to capture complementary aspects of message representation. An ablation study evaluating the impact of each loss component on clustering quality is shown in Table A6.

$$\begin{aligned} \mathcal{L}_{\text{total}} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{sign}}\mathcal{L}_{\text{sign}} \\ + \lambda_{\text{belief}}\mathcal{L}_{\text{belief}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} \end{aligned} \tag{5}$$

(1) Reconstruction Loss ($\mathcal{L}_{recon}$) preserves signed structural relationships. Predicted adjacency is:

$$\hat{A}_{ij} = \sigma(z_i^\top z_j), \tag{6}$$

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:

$$A_{\text{target}} = \frac{A_{\text{signed}} + 1}{2} \tag{7}$$

The reconstruction loss is:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} (\hat{A}_{ij} - A_{\text{target},ij})^2, \tag{8}$$

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

(2) Sign-Consistency Loss ($\mathcal{L}_{sign}$) regulates the persona subspace to respect social alignment. Positively linked pairs should have close embeddings; negatively linked pairs should be separated by margin M:

$$\begin{aligned} \mathcal{L}_{\text{sign}} = \frac{1}{|E^+|} \sum_{(i,j)\in E^+} \|z_{p,i} - z_{p,j}\|_2^2 \\ + \frac{1}{|E^-|} \sum_{(i,j)\in E^-} \max(0,\, M - \|z_{p,i} - z_{p,j}\|_2)^2 \end{aligned} \tag{9}$$

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

$$\mathcal{L}_{\text{belief}} = -\frac{1}{N}\sum_{i=1}^{N}\big[y_i \log \sigma(s_i) + (1 - y_i)\log(1 - \sigma(s_i))\big] \tag{10}$$

(4) Orthogonality Loss ($\mathcal{L}_{orth}$) enforces disentanglement by minimizing cross-covariance:

$$\mathcal{L}_{\text{orth}} = \left\| \frac{(z_b - \bar{z}_b)^\top (z_d - \bar{z}_d)}{N - 1} \right\|_F^2 \tag{11}$$

where $\| \cdot \|_F$ is the Frobenius norm (Golub and Van Loan, 2013) and $\bar{z}_b, \bar{z}_d$ are mean-centered embeddings.

4.3 Stage 2(b): Topic Modeling via Hierarchical Clustering

Having obtained the discourse embeddings $\{z_{d,i}\}_{i=1}^{N}$ from SiBeGNN, we identify discussion topics using hierarchical clustering. Each message t_i is represented by its discourse embedding $z_{d,i} \in \mathbb{R}^{d_d}$, which captures discourse characteristics independent of conspiratorial stance. To reduce dimensionality, we apply Principal Component Analysis (PCA) and retain the smallest number of components explaining $p\%$ of the variance. PCA helps remove noise while preserving the structure needed for clustering (Baligodugula and Amsaad, 2025; Jabari et al., 2024), and is commonly used before clustering methods such as K-Means and hierarchical clustering (Baranwal et al., 2025; Lötsch, 2024).

The reduced embeddings are clustered using agglomerative hierarchical clustering with Ward's linkage (Murtagh, 1983). The number of topics k^* is selected from $k \in [k_{\min}, k_{\max}]$ using the silhouette coefficient (Shahapure and Nicholas, 2020). To avoid redundant topics, we merge topic pairs whose centroid cosine similarity exceeds a threshold τ. We test multiple hyperparameter settings ($pca\_var \in \{0.5, 0.6, 0.7\}$, $merge\_th \in \{0.75, 0.8\}$, $k_{\min} \in \{2, 3, 4\}$, $k_{\max} = 20$). Across 18 configurations, coherence scores range from 0.360 to 0.386, indicating stable clustering. We report results using the best configuration in Table A5.

5 Results

Stage 1: Conspiratorial Content Classification: As shown in Table 3, RoBERTa-large achieved the highest performance (F1 = 0.866), while SEA-LION-v1-7B (AI Singapore, 2024) scored lowest (F1 = 0.653). We therefore use RoBERTa-large to label the full corpus of 553,648 messages.

Stage 2: Topic Modeling Results: Hierarchical clustering on the SiBeGNN discourse embeddings identifies seven discussion topics in the Telegram corpus (Table 2). These topics capture a range of conversations occurring in Singapore-based Telegram groups. The largest topics include *General Discussions* and *Banking and Finance*, which mainly contain everyday conversations such as shopping, food, financial advice, and casual chat. Other topics reflect more issue-focused discussions. For example, *Contradictions in Authority* and *Medical Concerns* contain discussions related to health policies and skepticism toward institutional decisions, while *General Legal Topics* focuses on practical legal questions and civic issues. Smaller sized topics such as *Group Moderation* and *Media Discussions* mainly include conversations about community rules and popular culture.

Linguistic analysis using LIWC (Boyd et al., 2022) further highlights differences between the topics. *General Legal Topics* uses more informational and neutral language, while *Medical Concerns* shows higher emotional and health-related language. *Media Discussions* contains informal conversational language related to entertainment and news. *Banking and Finance* includes analytical language focused on money and economic issues, whereas *Contradictions in Authority* contains higher levels of negative emotion and disagreement toward institutions. *Group Moderation* mainly contains short procedural messages about group rules.

Conspiratorial messages appear across multiple topics rather than concentrating within a single discussion cluster. Higher proportions occur in *Contradictions in Authority* and *Medical Concerns*, which often discuss institutional decisions and health policies. However, conspiratorial messages also appear in everyday topics such as *Banking and Finance* and *General Legal Topics*. This suggests that conspiratorial discourse is embedded within normal discussions rather than forming isolated communities.

We further evaluate clustering quality using topic coherence, silhouette score, and the Davies–Bouldin index, combined into a composite cDBI metric (Krasnov and Sen, 2019):

Topic Name	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 2: Topic modeling results with estimated conspiratorial message counts. Total estimated conspiratorial messages: 106,931.

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 3: 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.

$$cDBI = \frac{\text{Davies–Bouldin Index}}{\text{Average Coherence}} \tag{12}$$

Lower cDBI values indicate better clustering quality. Hierarchical clustering on SiBeGNN embeddings achieves the best performance (cDBI = 8.38), outperforming BERTopic (13.60) and other clustering approaches (33.13–67.27) as shown in Table 4. This indicates that the signed graph embeddings produce more coherent and better-separated topics. Ablation experiments further show that each component of the training loss contributes to clustering quality. Removing the orthogonality loss L_{orth}, sign-consistency loss L_{sign}, or belief-alignment loss L_{belief} increases cDBI by 58.2%, 94.6%, and 136.9%, respectively. Using only the reconstruction loss L_{recon} leads to the largest degradation (271.4%). These results show that modeling both message similarity and conspiratorial stance helps produce more meaningful topics.

Manual Topic Verification: To evaluate whether the discovered topics capture meaningful discussions, we compare them with nine expert-defined reference topics (Table 5). These topics are not used during model training; instead, they serve as a qualitative benchmark for assessing topic interpretability and coverage. Using expert-defined topics to interpret the outputs of unsupervised topic modeling methods is common in computational social science and discourse analysis (Grimmer and Stewart, 2013; Teufel, 2006; DiMaggio et al., 2013).

The nine reference topics were derived through qualitative analysis of a stratified sample of messages from the six Telegram groups. Two domain experts independently reviewed the messages to identify recurring discussion patterns, and a third expert resolved disagreements. The resulting topics include: *Technocratic Control vs Personal Autonomy*, *Betrayal by Trusted Institutions*, *Unequal Citizenship & Social Stratification*, *Elite Collusion & Loss of National Sovereignty*, *Information Manipulation & Manufactured Consent*, *Moral / Civilisational Decline*, *Awakening, Resistance & Minority Identity*, *Everyday Life Under Systemic Stress*, and *External Conflict as Moral Contrast*.

Table 5 compares different topic modeling methods based on how many of these nine reference topics appear in their discovered topics. Hierarchical clustering with SiBeGNN embeddings identifies seven reference topics while producing clear and interpretable topics. BERTopic with vanilla RoBERTa embeddings identifies five topics, while BERTopic with SiBeGNN embeddings identifies six topics. Other methods, including hierarchical clustering with vanilla RoBERTa embeddings, HDBScan, Gaussian mixture models, spectral clus-

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

Model	Ground Truth Overlap (/9)
Hierarchical clustering with SiBeGNN embeddings	7/9
Bertopic clusters with vanilla RoBERTa-large embeddings	5/9
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 topics coverage across clustering methods. Scores indicate the number of topics captured out of 9 possible ground truth topics

tering, and KMeans, identify fewer topics. Overall, hierarchical clustering using SiBeGNN embeddings provides the best balance between interpretability and topic coverage. The belief-aware embeddings help separate conspiratorial and non-conspiratorial discourse patterns while preserving semantic similarity, resulting in clearer and more coherent discussion topics.

6 Discussion and Implications

Methodological implications. Our signed belief graph–based clustering produces higher-quality topic clusters compared to standard methods. Hierarchical clustering on SiBeGNN embeddings achieves the lowest cDBI score (8.38) compared to other approaches (13.60–67.27), indicating more coherent and well-separated topics (Krasnov and Sen, 2019). The Sign-Disentanglement Loss helps separate conspiratorial stance from general discourse patterns, allowing the model to capture both semantic similarity and stance relationships between messages. This approach may be useful for other tasks involving polarized or stance-driven discussions, such as stance detection and political discourse analysis.

Platform governance and generalizability. The discovered topics show different levels of conspiratorial content. For example, *Medical Concerns* and *Contradictions in Authority* contain higher proportions of conspiratorial messages, while clusters such as *General Discussions* or *Group Moderation* contain fewer. These patterns suggest that moderation strategies may benefit from topic-aware approaches rather than treating conspiratorial content as a single homogeneous category. While our framework is general and can be applied to other platforms, the specific topics identified in this study reflect the socio-political context of Singapore. Future work should examine whether similar patterns appear in other regions and platforms.

7 Conclusion

We propose a two-stage framework that combines transformer-based classification with Signed Belief Graph Neural Networks (SiBeGNN) to perform topic modeling on conspiratorial discussions. The approach learns belief-aware message embeddings and identifies seven discussion topics across 553,648 Telegram messages. These topics reveal that conspiratorial content often appears alongside routine social interactions rather than as a separate discourse space. Methodologically, the Sign-Disentanglement Loss helps separate conspiratorial stance from general discourse characteristics, leading to more coherent and interpretable topic clusters (cDBI = 8.38). This framework provides a practical approach for studying belief-driven discourse on online platforms and may be useful for related tasks such as stance detection, misinformation analysis, and political discourse modeling.

8 Limitations

This study has several limitations that suggest directions for future work. First, our analysis relies on embeddings from a single transformer model. Different language models or multimodal representations may capture additional patterns in conspiratorial discussions. Second, some design choices in our pipeline, such as normalization and similarity

thresholds, are heuristic and may be further optimized for specific datasets or platforms. Another limitation is that our analysis relies only on textual features. Online discussions are shaped not only by message content but also by behavioral and interaction patterns within communities. For example, users who primarily read messages without posting (often referred to as *lurkers*) remain invisible in text-based analyses but may still influence how information spreads. Future work could incorporate additional signals such as posting frequency, temporal activity patterns, engagement metrics, response latency, and network position to better capture these dynamics. Finally, our dataset consists of Telegram groups based in Singapore, which reflects a specific socio-political and cultural context. The discussion topics identified in this study may therefore differ from those observed on other platforms or in other regions. Future research could study topic evolution over time (Shimgekar et al., 2025b,a), incorporate interaction networks, and compare results across platforms and geographic contexts.

9 Ethical Considerations

We implemented several safeguards to protect user privacy, including anonymization, paraphrasing of example messages, and analysis at an aggregate level. Although the dataset contains publicly accessible Telegram messages, care was taken to avoid exposing identifiable information. We also recognize that labeling messages as conspiratorial can be sensitive. Institutional skepticism and criticism of policies are common in public discourse and should not automatically be treated as misinformation. Our annotation guidelines therefore focused on identifying claims that attribute events to hidden or coordinated actors without supporting evidence, rather than general criticism or personal opinions. All annotation was performed by domain experts involved in the project. Messages were labeled based only on their textual content without attempting to infer author identity or intent. Annotators were informed that the dataset may contain offensive or sensitive language typical of online political discussions. Finally, while we plan to release the methodology and code, access to the raw dataset will remain restricted to comply with institutional ethical guidelines and to reduce the risk of misuse.

10 AI Involvement Disclosure

AI-assisted language editing was used exclusively to improve grammar and readability. The study design, analyses, interpretations, and experiments were conducted fully by the authors.

References

Ahmed Mohammad Abdou. 2021. Good governance and covid-19: The digital bureaucracy to response the pandemic (singapore as a model). *Journal of Public Affairs*, 21(4):e2656.

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, Irene Corcuera-Platas, Juan Fernando Sánchez-Rada, and Carlos A. Iglesias. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246.

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.

V. V. Baligodugula and F. Amsaad. 2025. Unsupervised learning: Comparative analysis of clustering techniques on high-dimensional data. *arXiv preprint arXiv:2503.23215*.

S. Baranwal and 1 others. 2025. Optimizing clustering of cdr3 sequences using nlp embeddings, pca, and k-means. *Frontiers in Bioinformatics*.

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. 2023. 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.

Keyu Chen, Ashley Feng, Rohan Aanegola, Koustuv Saha, Allie Wong, Zach Schwitzky, Roy Ka-Wei Lee, Robin O'Hanlon, Munmun De Choudhury, Frederick L Altice, and 1 others. 2022. Categorizing memes about the ukraine conflict. *International Conference on Computational Data and Social Networks*, pages 27–38.

Gang Cheng, Qinliang You, Lei Shi, Zhenxue Wang, Jia Luo, and Tianbin Li. 2023. A survey of topic models: From a whole-cycle perspective. *Journal of Intelligent & Fuzzy Systems*, 45(6):9929–9953.

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.

Eunsol Choi, Hannah Rashkin, Luke Zettlemoyer, and Yejin Choi. 2016. Document-level sentiment inference with social, faction, and discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 333–343.

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.

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606.

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

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.

Gene H Golub and Charles F Van Loan. 2013. *Matrix computations*. JHU press.

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*.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

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

Mohamad Hoseini, Philipe Melo, Fabrício Benevenuto, Anja Feldmann, and Savvas Zannettou. 2021. On the globalization of the qanon conspiracy theory through telegram. *arXiv preprint arXiv:2105.13020*.

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

I. K. O. Jabari and 1 others. 2024. Learning-augmented k-means clustering using dimensional reduction. *arXiv preprint arXiv:2401.03198*.

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*.

Jörn Lötsch. 2024. Comparative assessment of projection and clustering combinations. *ScienceDirect*.

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. In *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*, pages 3–18. Walter de Gruyter.

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*.

Lynnette Hui Xian Ng and Jia Yuan Loke. 2020. Analyzing public opinion and misinformation in a covid-19 telegram group chat. *IEEE Internet Computing*, 25(2):84–91.

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*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Markus Reiter-Haas, Beate Klösch, Markus Hadler, and Elisabeth Lex. 2024. Framing analysis of health-related narratives: Conspiracy versus mainstream media. *arXiv preprint arXiv:2401.10030*.

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.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

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.

Simone Teufel. 2006. Argumentative zoning for improved citation indexing.

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.

Tran Hien Van, Abhay Goyal, Muhammad Siddique, Lam Yin Cheung, Nimay Parekh, Jonathan Y Huang, Keri McCrickerd, Edson C Tandoc Jr, Gerard Chung, and Navin Kumar. 2023. How is fatherhood framed online in singapore? *arXiv preprint arXiv:2307.04053*.

Teun A. van Dijk. 1998. *Ideology: A Multidisciplinary Approach*. Sage Publications, London.

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

Hanlin Xiao, Mauricio A Álvarez, and Rainer Breitling. 2026. Disentangling similarity and relatedness in topic models. *arXiv preprint arXiv:2603.10619*.

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 [...].

Dimension	Conspiratorial (Label = 1)	Non-Conspiratorial (Label = 0)
Core Definition	Message presents an explanatory narrative attributing events or conditions to hidden, coordinated actors pursuing malevolent or deceptive goals that contradict official or mainstream explanations.	Message discusses events, policies, or institutions without attributing them to secret coordination or hidden malicious intent.
Hidden Actors	Explicit or implicit reference to secretive groups (e.g., "elites," "they," "globalists," "the system") operating covertly.	No hidden agents are implied; actors are named institutions, policies, individuals, or abstract systems without secrecy claims.
Coordination	Actors are described as intentionally coordinating actions behind the scenes (e.g., planning, engineering, orchestrating outcomes).	No claim of covert coordination; actions may be attributed to incompetence, bureaucracy, coincidence, or structural issues.
Intent / Goal	Attributes malevolent intent such as harm, control, deception, exploitation, or suppression of truth.	Does not assign malicious intent; may express dissatisfaction, concern, or critique without hidden agendas.
Epistemic Stance	Rejects official explanations by asserting that the "real truth" is concealed; often framed as exclusive or privileged knowledge.	May question or critique official accounts but remains open to evidence, uncertainty, or multiple explanations.
Narrative Structure	Contains a causal narrative linking hidden actors $\rightarrow$ secret actions $\rightarrow$ observed outcomes.	Lacks a covert causal narrative; statements may be descriptive, interrogative, or experiential.
Use of Evidence	Relies on insinuation, pattern-matching, or distrust rather than verifiable evidence; evidence may be framed as "suppressed" or "censored."	References verifiable information, personal experience, or openly available sources without alleging suppression.
Examples of Inclusion	"They are intentionally poisoning people through vaccines to control the population."	"I don't trust this vaccine because it was developed quickly and the side effects worry me."
Examples of Exclusion	—	Policy critique, satire, personal anecdotes, logistical complaints, or emotional reactions without hidden-agent attribution.
Ambiguous Cases	If hidden actors are vaguely referenced without clear coordination or intent, message is discussed in calibration and labeled conservatively.	If no clear covert narrative is present, the default label is non-conspiratorial.

Table A2: Operational criteria distinguishing conspiratorial and non-conspiratorial content used for binary labeling.

Table A3: Hyperparameter ablation results showing accuracy and F1 performance.

Learning Rate	Batch Size	Epochs	Dropout	Max Length	Accuracy	F1
2e-5	16	5	0.1	256	0.85	0.87
1e-5	16	5	0.1	256	0.83	0.85
3e-5	16	5	0.1	256	0.84	0.85
5e-5	16	5	0.1	256	0.81	0.83
2e-5	8	5	0.1	256	0.84	0.86
2e-5	32	5	0.1	256	0.84	0.85
2e-5	16	3	0.1	256	0.83	0.85
2e-5	16	7	0.1	256	0.84	0.86
2e-5	16	5	0.0	256	0.83	0.85
2e-5	16	5	0.2	256	0.84	0.86
2e-5	16	5	0.1	128	0.83	0.84
2e-5	16	5	0.1	512	0.85	0.87

Table A4: Summary of discourse-level and semantic features used for modeling belief expression in text.

Feature	Discourse Dimension Captured	Computation Method	Lexicon / Model Used	Output Format	Example Text
Epistemic Modality	Degree of epistemic commitment (certainty vs. hedging)	Lexicon-based contrast between certainty and hedging expressions	Certainty markers: {definitely, clearly, certainly, undeniably, for sure} Hedging markers: {maybe, perhaps, I think, not sure, possibly}	Scalar $\in (-1, 1)$; positive = high certainty, negative = hedging	"I think this is probably true, but I'm not sure the authorities clearly understand what's happening."
Agency	Attribution of responsibility and intentionality	Lexicon-based contrast between action-oriented active phrases and passive or impersonal constructions	Active agency phrases: {I will, we must, we can, take action} Passive / impersonal phrases: {it is said, was done, is believed}	Scalar $\in (-1, 1)$; positive = explicit agency, negative = obscured agency	"We must act now and hold them accountable, not just accept what is said on the news."
Sentiment Polarity	Evaluative stance toward propositions	Transformer-based sentiment classification	Pretrained model: `cardiffnlp/twitter-roberta-base-sentiment-latest`	Continuous scalar $\in [-1, 1]$; positive = favorable, negative = unfavorable	"This policy is terrible and has caused nothing but harm to ordinary people."
Emotion Spectra	Affective intensity and emotional framing	Transformer-based emotion classification; probabilities extracted for core emotions	Pretrained model: `j-hartmann/emotion-english-distilroberta-base` Emotions: {joy, anger, fear, sadness}	4-dimensional vector of emotion probabilities	"I'm scared about where this is heading, and honestly it makes me angry that no one is listening."
Semantic Content (baseline)	Propositional meaning of the message	Mean-pooled contextual embeddings	Pretrained model: `roberta-large`	Dense vector (768-dimensional)	"The government announced new regulations affecting online platforms starting next month."

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 A5: 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 A6: 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.

Table A7: Approximate Topic Clusters Using Standard Embeddings and Topic Clustering

Topic Cluster (Standard Embeddings)	Description	Common Keywords	# Messages	Avg. Length	# Conspiratorial	% Conspiratorial
Legal & Policy Discussions	Laws, regulations, mandates, civic rules	law, policy, rules, mandate, legal	72,000	155	14,900	20.7%
Health & Medicine	Vaccines, symptoms, hospitals, health advice	vaccine, medical, hospital, covid, health	60,500	165	13,200	21.8%
Finance & Economy	Banking, inflation, jobs, cost of living	bank, money, rates, economy, cost	75,200	225	13,100	17.4%
News & Media	News reporting, pop culture, viral stories	news, media, video, concert, viral	68,000	285	11,900	17.5%
Government & Authority	Governance, enforcement, public institutions	government, authority, rules, enforcement	55,000	140	12,400	22.5%
Community & Moderation	Group rules, admin actions, etiquette	admin, rules, ban, group, respect	6,000	40	1,096	18.3%
Casual & Everyday Chat	Shopping, food, travel, daily life	shopping, food, travel, chat, daily	216,448	120	40,531	18.7%

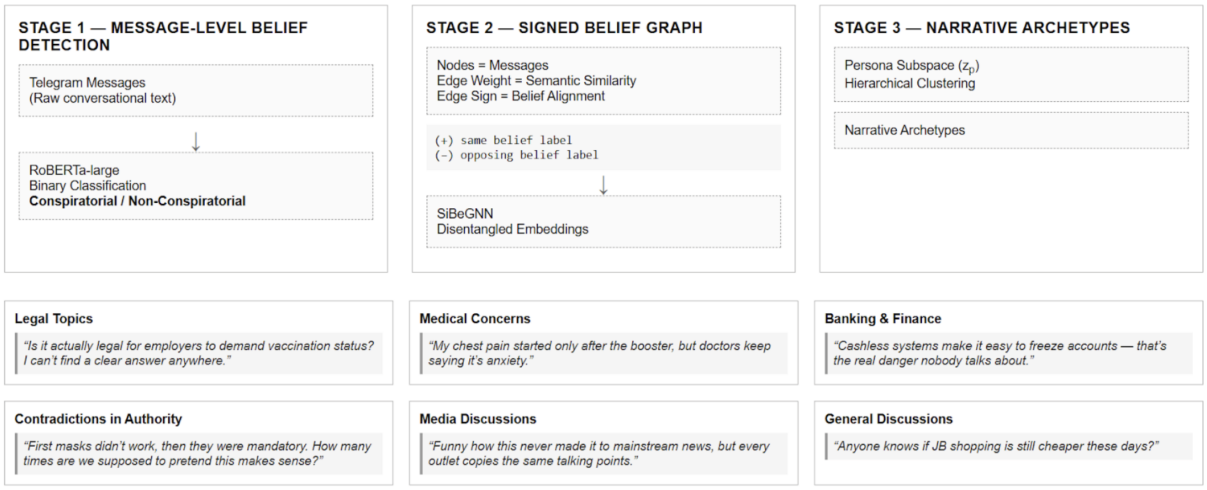

Figure A1: Overall architecture of the proposed framework, illustrating discourse-level feature extraction, persona disentanglement, and downstream clustering.