Large Language Model-Guided Disentangled Belief Representation Learning on Polarized Social Graphs

Jinning Li, Ruipeng Han, Chenkai Sun, Dachun Sun, Ruijie Wang, Jingying Zeng†, Yuchen Yan, Hanghang Tong, Tarek Abdelzaher
Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
†College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
{jinning4,ruipeng2,chenkai5,dsun18,ruijiew2,yucheny5,htong,zaher}@illinois.edu, joyzeng@gatech.edu†

***Abstract*—The paper advances *belief representation learning in polarized networks* – the mapping of social beliefs espoused by users and posts in a polarized network into a disentangled latent space that separates (the members and beliefs of) each side. Our prior work embeds social interaction data, using non-negative variational graph auto-encoders, into a disentangled latent space. However, the interaction graphs alone may not adequately reflect similarity and/or disparity in beliefs, especially for those graphs with sparsity and outlier issues. In this paper, we investigate the impact of limited guidance from Large Language Models (LLMs) on the accuracy of belief separation. Specifically, we integrate social graphs with LLM-based soft labels as a novel weakly-supervised interpretable graph representation learning framework. This framework combines the strengths of graph- and text-based information, and is shown to maintain the interpretability of learned representations, where different axes in the latent space denote association with different sides of the divide. An evaluation on six real-world Twitter datasets illustrates the effectiveness of the proposed model at solving stance detection problems, demonstrating 5.9%-6.5% improvements in the accuracy, F1 score, and purity metrics, without introducing a significant computational overhead. An ablation study is also discussed to study the impact of different components of the proposed architecture.**

***Index Terms*—Large Language Models, Weak Supervision, Graph Auto-Encoders, Interpretability, Social Networks**

I. Introduction

Beliefs posted on social networks are propagated through interactions among users and messages [1]. Understanding and modeling these beliefs helps researchers explore the opinions, preferences, and evolution of individuals and communities. Mainstream methods represent social stances [2], [3] or beliefs [4], [5] as vectors (embeddings), enabling analysts to discover social facts and trends by studying the positions, similarities, and dynamics of belief representations. Additionally, belief representations facilitate various downstream tasks such as polarization detection [6] and stance prediction [7].

Previous belief representation learning work has utilized techniques such as graph neural networks (GCN) [8] and variational graph auto-encoders (VGAE) [9] to encode user and message embeddings. An interpretable belief representation is proposed in [4], where users and messages are mapped into a disentangled belief embedding space, with each axis aligned with one side of an ideological divide. The coordinates of a point on each axis in the latent space represent the strength of its affiliation with the corresponding side's ideology. However, the efficacy of interpretable separation of polarized communities based on their interaction patterns is limited by complexities found in real-world data and challenges arising from the sparsity of observations. For example, members on opposite sides of a politically polarized network may repost some of the same content (e.g., soccer game results, as opposed to political views). Conversely, members of the same side might sometimes have no interaction due to differences in interests. Furthermore, spam accounts may exhibit abnormal behaviors such as reposting random messages from both sides. Thus, real-world social interaction graphs may not fully adhere to the homophily assumption, where users prefer to interact with similar entities [10].

To address the limitations of inferring polarization from interaction patterns only, recent research has also explored representations of actors and messages based on the textual data they post [11], [12]. The recent advances in natural language processing, including large language models (LLMs), have notably improved text-based representation learning, particularly in zero-shot or few-shot scenarios. However, social messages with sarcasm or abbreviations are often less informative and may prove challenging to interpret in the absence of sufficient context, resulting in sub-optimal representations.

To address these issues, studies have investigated combining social graphs and message text [13], including enhancing graph embeddings with LLM encoder features [14], [15] or incorporating graph structure as additional inputs to LLM encoders [16]–[18]. Most of these efforts focus on embedding-based integration, utilizing message embeddings as node features [14], [19], or to smooth the adjacency matrix based on cosine similarity [20]. These approaches combine the benefits of graph and text modalities to create composite node representations in social interaction graphs. However, the need to compute a text embedding for *every* post in these approaches increases their computation overhead.

To reduce the computational overhead of belief representation learning approaches that jointly exploit text and graph structure, instead of embedding every post, in [21], the authors propose to utilize labels generated by LLM decoders to distantly supervise the training of a graph model, which is lightweight and effective for node classification tasks. However, the approach does not ensure the *interpretability* of learned representations from the integration of LLM decoder

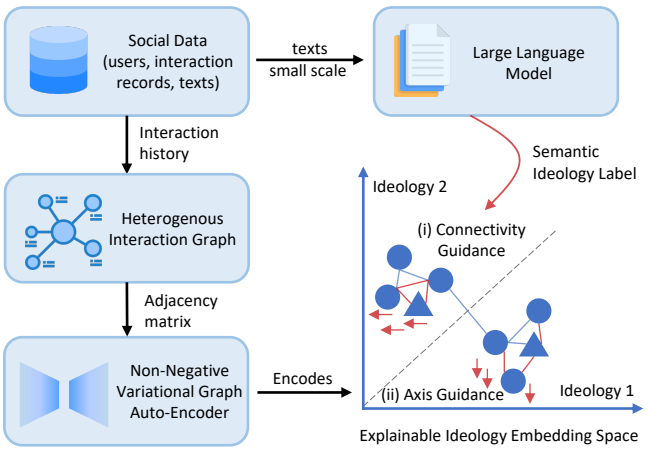

Fig. 1: The proposed SGVGAE model framework processes input comprising social interaction records and message texts. A Non-Negative VGAE maps user and message nodes to an interpretable belief embedding space, where we further refine the embedding using guidance from semantic belief labels generated by LLMs. Techniques such as connectivity guidance and axis guidance enhance the interpretability and robustness of the model in the face of uncertain graph topology with a limited number of semantic labels.

labels with the graph model.

In this paper, we propose a Semantic-Guided Variational Graph Auto-Encoder (**SGVGAE**) model to enhance *interpretable belief representation learning* under the weak supervision of LLM-generated belief labels. The approach is the first to jointly (i) utilize both graph and text embeddings, (ii) ensure efficiency by labeling only a small subset of messages, and (iii) retain interpretability of the latent space by maintaining an association between axes and group ideologies on different sides of an ideological divide. Other solutions attain at most two out of the above three desirable properties.

SGVGAE proposes two semantic guidance techniques to improve the quality of learned representations while preserving the interpretability of the belief embedding space: (i) *Connectivity Guidance* increases the connectivity among messages with the same semantic belief (LLM-generated soft) label, and (ii) *Axis Guidance* enhances the alignment of users or messages with the correct belief axis while penalizing alignment with incorrect axes. Additionally, we introduce a learnable gate that allows the model to selectively leverage information from the initial graph structure or the belief labels. The weak supervision provided by the soft LLM labels enhances the robustness of the SGVGAE model, particularly when the social interaction graph is sparse or unreliable. The approach improves the quality of learned representations and expands the range of downstream applications. Our experiments demonstrate that associating as little as 5% of the messages with LLM-generated soft labels can significantly improve the performance and stability of SGVGAE.

We evaluate the proposed SGVGAE model on six real-world Twitter datasets for belief detection. We compare the accuracy, macro F1 score, and purity score of SGVGAE with graph-based models (e.g., VGAE and its variants), text-based models (e.g., GPT-3.5 and GPT-4), and multi-modal models (e.g. TIMME [14]) in zero-shot, few-shot, and semi-supervised fine-tuning scenarios. The proposed SGVGAE model outperforms the best baseline by 5.97% in accuracy, 6.06% in macro F1 score, and 6.51% in purity score. We also explore the performance of SGVGAE under different scales of semantic guidance. An ablation study is conducted to validate the effectiveness of the proposed semantic guidance techniques.

II. Problem Statement

The input for belief representation learning consists of typical social network data, including entities and interactions. Entities comprise users $U = \{u_1, u_2, ..., u_n\}$ and messages $M = \{v_1, v_2, ..., v_m\}$. Messages can be tweets on Twitter, posts on Reddit, etc. Note that we only consider deduplicated identical messages as v, meaning that, for example, all retweets of the same original tweet are treated as the same message. For each message v, the data also includes a corresponding textual context t, such as the content of the tweet. Interaction records I are relational triplets between users and messages, denoted by $(u, r, v) \in I$, where r represents the type of interaction, such as *retweet* or *post* on Twitter.

The objective of belief representation learning is to predict the belief label l (such as *conservative* versus *liberal* in political polarization or such as *sect A* versus *sect B* in an arbitrary divide between two groups) for each user and message, as well as their belief strengths for every belief category. Typically, the belief strengths are represented as an embedding (vector) $\boldsymbol{z} \in Z$ for all of n users or m messages, i.e., $|Z| = n + m$. Assuming there are d pre-defined belief categories in the problem, $\boldsymbol{z} \in \mathbb{R}^d$ is a d-dimensional embedding. In this paper, we treat belief representation learning as a *weakly-supervised learning* or *distantly-supervised learning* task, where a small scale of soft belief labels generated by LLMs or pre-trained models are used as additional input.

III. Semantic-Guided Non-Negative Variational Graph Auto-Encoders

In this section, we present the formulation of the Semantic-Guided Variational Graph Auto-Encoders (SGVGAE) model, which encompasses (i) Non-Negative Variational Auto-Encoders, (ii) LLM-based Semantic Labeling of beliefs, (iii) Connectivity Guidance, and (iv) Axis Guidance techniques.

The proposed SGVGAE model represents interactions on social media by a heterogeneous undirected bipartite graph. The graph nodes consist of users and messages, resulting in a total of n+m nodes. For simplicity, we assume the first n nodes represent users and the subsequent m nodes represent messages. The graph edges are formulated as a weighted adjacency matrix $\boldsymbol{A}^r \in \mathbb{R}^{(n+m)\times(n+m)}$, where r denotes the edge type. The graph edges are connected with the interaction triplets between users and messages, with $\boldsymbol{A}^r_{i,n+j} = \boldsymbol{A}^r_{n+j,i} = 1$ if $(u_i, r, v_j) \in I$ and $\boldsymbol{A}^r_{i,j} = 0$ otherwise. In this paper, we simplify the multi-relation adjacency matrix to a general adjacency

matrix by taking the element-wise maximum $\boldsymbol{A} = \max_r(\boldsymbol{A}^r)$, considering that all relations (e.g., retweet, post) are positive (consentaneous). The resulting graph is bipartite, meaning that edges only exist between users and messages, i.e., $\boldsymbol{A}_{i,j} \equiv 0$ for $1 \leq i, j \leq n$ or $i, j \geq n+1$. While the bipartite graph is a straightforward model of input interactions, in Section III-D, we will discuss how to introduce message-message edges with guidance from LLMs to enhance the performance.

A. Inference Model (Encoder)

We adopt the non-negative inference model introduced in [4] to create a disentangled orthogonal latent space. The inference model encodes the general-relation adjacency matrix $\boldsymbol{A}$ as the belief embedding $\boldsymbol{Z} \in \mathbb{R}^{(n+m)\times d}$. In this paper, we use the identity matrix as the node feature $\boldsymbol{X} \in \mathbb{R}^{(n+m)\times(n+m)}$ to capture the topological information. We apply symmetrical normalization based on the degree of nodes to the adjacency matrix $\bar{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$, where $\boldsymbol{D}$ is a diagonal matrix of degrees, with $D_{i,i} = \sum_{j=1}^{n+m} D_{i,j}$.

We utilize an L-layer Graph Convolutional Network (GCN) as the backbone of the encoder. Let $\boldsymbol{C}^k$ represent the k-th layer graph convolution and initialize $\boldsymbol{C}^1 = \boldsymbol{X}$; the first $L-1$ layers of the encoder network can be formulated as

$$\boldsymbol{C}^k = \gamma(\bar{\boldsymbol{A}}\boldsymbol{C}^{k-1}\boldsymbol{W}^k), 2 \leq k \leq L-1, \tag{1}$$

where γ denotes the activation function. We assume the belief embedding follows an element-wise rectified Gaussian Distribution $\mathcal{N}_+$. In the L-th output layer of the encoder, we compute the means and standard deviations as

$$\boldsymbol{\mu} = ReLU(\bar{\boldsymbol{A}}\boldsymbol{C}^{L-1}\boldsymbol{W}^{\boldsymbol{\mu}}),\ \boldsymbol{\delta} = ReLU(\bar{\boldsymbol{A}}\boldsymbol{C}^{L-1}\boldsymbol{W}^{\boldsymbol{\delta}}), \tag{2}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ represent the mean and standard deviation, respectively. Using the reparameterization trick, we can sample from $\mathcal{N}_+(\boldsymbol{\mu}, \boldsymbol{\delta})$ to compute the belief embedding $\boldsymbol{Z}$.

B. Generative Model (Decoder)

We design the decoder using an inner product to maintain the orthogonality of the belief embedding space. The decoder calculates the probability of the reconstructed adjacency matrix $\boldsymbol{A}_{i,j}$ given the belief embeddings $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ as the sigmoid probability of their inner product, which is formulated as

$$p(\boldsymbol{A}_{i,j}|\boldsymbol{z}_i, \boldsymbol{z}_j) = \sigma(\boldsymbol{z}_i^T \boldsymbol{z}_j), \tag{3}$$

where σ is the sigmoid function. Practically, we model the reconstruction loss $\mathcal{L}_{rc}$ as the element-wise cross-entropy loss,

$$\mathcal{L}_{rc} = -\sum_{i=0,j=0}^{n+m}\Big[\omega_p^{\boldsymbol{A}}\boldsymbol{A}_{i,j}\log p(\boldsymbol{A}_{i,j}|\boldsymbol{z}_i, \boldsymbol{z}_j) + (1-\boldsymbol{A}_{i,j})\log(1-p(\boldsymbol{A}_{i,j}|\boldsymbol{z}_i, \boldsymbol{z}_j))\Big], \tag{4}$$

where $\omega_p^{\boldsymbol{A}}$ is the positive-class weight used to balance the number of positive and negative samples. It is calculated as the ratio of the negative count to the positive count, denoted as $\omega_p^{\boldsymbol{A}} = \Big[\sum_{i,j}(1-\boldsymbol{A}_{i,j})\Big] / \sum_{i,j}\boldsymbol{A}_{i,j}$. By optimizing the reconstruction loss $\mathcal{L}_{rc}$, the non-negative VGAE maps all actors and messages to a non-negative embedding space. The orthogonality of the embedding space is ensured by the non-negative encoder and the inner-product decoder, resulting in a disentangled representation. An illustrative example of the interpretable belief embedding space is shown in Figure 1, where each axis represents a belief, and the corresponding coordinate indicates the belief strength. Typically, unsupervised graph embedding learning models require an additional clustering algorithm to separate different beliefs. However, with the interpretable embedding space, one advantage is that further clustering algorithms are not needed to separate different beliefs. The predicted belief labels can be assigned simply by taking $\arg\max$ over the embeddings.

The VGAE and non-negative VGAE models rely on the topology of the graph, making them potentially vulnerable to sparsity or uncertainty in the graph structure. Belief embedding learning essentially aims to find the optimal split of the graph by mapping nodes to appropriate positions in the embedding space. However, when the dataset is not highly polarized, graph-based models like VGAEs can only find sub-optimal splits without additional information or guidance. For instance, a real-world social interaction network may have many small isolated components that are weakly connected to the main component (or two main components in case of polarization). While VGAE models may effectively separate the two main components (clusters) in a polarized network, they struggle to optimally align other isolated components with the right axis and struggle when some nodes (e.g., neutral individuals or spam bots interacting with both sides of the ideological divide), leading to an improper mapping. Additionally, the general sparsity of social networks significantly impacts the performance of graph-based models. A very sparse graph without additional guidance may result in suboptimal behaviors.

One possible solution is to incorporate information from the text. The existing research has explored encoding message text with large language model encoders [12], [22]–[24]. The encoded message embeddings are typically used as node features [14], [19] for graph convolutional networks or to compute message-message cosine similarity to smooth the adjacency matrix [20]. However, these existing methods may have limitations including: (i) Textual embeddings from pre-trained language encoders often capture token distributions rather than the desired belief information. (ii) Using message-message cosine similarity to smooth the adjacency matrix introduces difficulties in choosing thresholds for controlling sparsity. (iii) Applying message embeddings to all node features or message-message edges introduces a significant overhead, as shown in the experiments of Section IV-F (iv) Finally, the used optimization objectives are often not designed to be compatible with the interpretability.

To incorporate text information with graphs and maintain the interpretability, the proposed SGVGAE model (i) utilizes decoder-based generative LLMs like GPT-3.5 and designs a prompt with demonstrative belief-prediction examples to more precisely infer the *semantic belief labels* for a small

subset of messages. (ii) We introduce *connectivity guidance* that uses semantic belief labels to increase intra-belief node connectivity. (iii) We propose *axis guidance* that introduces an additional soft constraint for semantically labeled messages to reside in the correct belief axis, which also guides the split of nodes with different beliefs. (iv) While the proposed soft constraints are compatible with the disentangled embedding space, we further propose normalization and gated fusion to allow the SGVGAE model to selectively balance the usage between graph and textual signals, enhancing its robustness. The details are elaborated in the following sections.

C. Semantic Labeling of Belief

We design and utilize a prompt with demonstrative examples for decoder-based LLMs to the inference message belief, i.e. we instruct the language model to return the belief label l for each message v. For example, the prompt for the Philippines Enhanced Defense Cooperation Agreement (EDCA) dataset is shown in Figure 2, where some repeated patterns are omitted for brevity. Empirically, we find providing additional demonstrative examples, as well as dataset-specific contexts, can facilitate LLMs in understanding the definitions of *sides* in the local ideological divide, which therefore produces more reliable soft labels. While human-annotated labels are also compatible with the proposed weak supervision framework, the LLM-based soft labeling has the advantage of automating the whole pipeline and reducing the annotation cost.

```
Given a set of tweets, determine if each of them
    expresses a Pro-EDCA (liberal) or Anti-EDCA
    (conservative) belief about the Enhanced
   Defense Cooperation Agreement (EDCA). In
   very rare cases there will be neutral views.
Here are some examples:
1: RT @JoeBiden: On this International Day of
   People with Disabilities, we'll be fully
   commit..." This tweet on the US election
   dataset expresses liberal belief because it
   expresses care for disabilities.
2: ...
The expected response format is an array of JSON
    objects as follows: ...
Inference the belief of the following tweets:
1: ...
```

Fig. 2: An example of an LLM prompt used as context for soft labeling of two sides in Philippines polarization analysis regarding Enhance Defense Cooperation Agreement.

D. Connectivity Guidance

Leveraging the semantic belief labels L, we enhance intra-belief connectivity, as illustrated by the red edges in Figure 1. Connectivity guidance is expected to result in a clearer separation of different belief groups. This technique also allows adjusting the scale of semantic belief labels to apply for the connectivity guidance. Suppose we want to apply k belief labels $L = \{l_1, l_2, ..., l_k\}$ to the corresponding messages $V = \{v_1, v_2, ..., v_k\}$. We integrate the intra-belief edges of belief labels into the original adjacency matrix $\boldsymbol{A}$ to construct an enriched adjacency matrix $\boldsymbol{A}^*$, which is formulated as,

$$\begin{aligned} \boldsymbol{A}^*_{i,j} = \boldsymbol{A}^*_{j,i} = g_{(i,j)},\ g_{(i,j)} \in [0,1] \\ \text{iff } (i,j \geq n+1) \wedge (v_i, v_j \in V) \wedge (l_i = l_j), \\ \boldsymbol{A}^*_{i,j} = \boldsymbol{A}_{i,j} \text{ otherwise,} \end{aligned} \tag{5}$$

where $i, j \geq n+1$ indicates that indices i, j represent messages. $v_i, v_j \in V$ denotes that we want to apply connectivity guidance between message v_i, v_j, and $l_i = l_j$ signifies that they share the same belief label. We also propose a one-init gated fusion utilizing the *trainable* gate parameter $g_{(i,j)}$. Practically, this trainable gate parameter can be obtained by applying the sigmoid function $\sigma(\cdot)$ to a trainable floating-point parameter w_g, which is infinitely initialized as $w_g := +\infty$. This ensures $g_{(i,j)} = \sigma(w_g) = 1$ in the first epoch, meaning that by default, full connectivity guidance is applied to the adjacency matrix. This gated fusion allows the model training process to selectively choose information from graph topology or textual guidance. Based on Equation 5, we update the reconstruction loss function with the connectivity guidance as

$$\begin{aligned} \mathcal{L}^*_{rc} = -\sum_{i=0,j=0}^{n+m} \Big[\omega_p^{\boldsymbol{A}^*} \boldsymbol{A}^*_{i,j} \log p(\boldsymbol{A}^*_{i,j} | \boldsymbol{z}_i, \boldsymbol{z}_j) \\ + (1 - \boldsymbol{A}^*_{i,j}) \log(1 - p(\boldsymbol{A}^*_{i,j} | \boldsymbol{z}_i, \boldsymbol{z}_j)) \Big], \end{aligned} \tag{6}$$

E. Axis Guidance

While connectivity guidance effectively enhances the separation of nodes by attracting intra-belief nodes and repelling nodes across beliefs, it does not enhance the alignment with the belief axis. Therefore, we introduce the axis guidance to enhance the alignment of node embeddings with the belief axis. Assume there are d belief categories (and therefore d axes); we denote the unit vectors (d-dimensional) for each axis as $\boldsymbol{e}$. Therefore, we have $\langle \boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_d \rangle = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

Similar to Section III-D, suppose we want to apply k semantic belief labels $L = \langle l_1, l_2, ..., l_k \rangle$. The corresponding belief embeddings are $\boldsymbol{Z} = \langle \boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_k \rangle$. We denote a predefined mapping function $f(\cdot) : k \to d$ that maps a label l to the unit vector $\boldsymbol{e}$ of the corresponding axis. We then retrieve the matched axis unit vectors for the k semantic belief labels, which can be denoted by $\boldsymbol{E} = \langle f(l_1), f(l_2), ..., f(l_k) \rangle = \langle \boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_k \rangle$. Note that duplicates may exist in $\boldsymbol{E}$ because multiple messages can be simultaneously aligned with the same belief axis.

We construct an alignment matrix $\boldsymbol{M} \in \mathbb{R}^{(k)\times(k)}$ as the ground-truth, for which we have

$$\boldsymbol{M}_{i,j} = 1 \text{ iff } l_i = l_j,\ \boldsymbol{M}_{i,j} = 0 \text{ otherwise, } \forall i, j. \tag{7}$$

Similar to the loss function for the decoder in Equation 4, our objective is to maximize the likelihood $p(\boldsymbol{M} | \boldsymbol{Z}, \boldsymbol{E})$, where $\boldsymbol{Z}, \boldsymbol{E} \in \mathbb{R}^{k \times d}$. To ensure compatibility with the graph reconstruction loss $\mathcal{L}^*_{rc}$ and maintain the interpretability of

TABLE I: Dataset Statistics. The number of annotations represents the number of manually labeled tweets for evaluation, which will not be selected for semantic guidance or semi-supervision. The average degree represents the overall sparsity.

Dataset	Keywords	#Tweet	#User	Avg. Degree	#Annotation	Label Types
Crime	crime	2321	451	5.81	659	Pro/Anti Government
EDCA	edca	5220	912	7.13	877	Pro/Anti EDCA
Energy China	energy, china	5951	3047	3.67	845	Pro/Anti China
US Military Philippine	military, philippine, south china sea	5757	3475	3.01	1627	Pro/Anti Western
Labor and Migration China	labor, migration, china	2719	1563	3.18	618	Pro/Anti China
Social and Economic Issues	social, economy	15902	2326	10.59	6179	Pro/Anti Philippines

the embedding space, we adopt the similar inner product and sigmoid function to compute the likelihood and loss function:

$$p(\boldsymbol{M}_{i,j}|\boldsymbol{z}_i, \boldsymbol{e}_j) = \sigma(\boldsymbol{z}_i^T \boldsymbol{e}_j), \tag{8}$$

based on which the loss function for axis guidance can be formulated as,

$$\mathcal{L}_{axis} = -\sum_{i=0,j=0}^{k} \Big[\omega_p^{\boldsymbol{M}} \boldsymbol{M}_{i,j} \log p(\boldsymbol{M}_{i,j}|\boldsymbol{z}_i, \boldsymbol{e}_j) + (1-\boldsymbol{M}_{i,j}) \log(1-p(\boldsymbol{M}_{i,j}|\boldsymbol{z}_i, \boldsymbol{e}_j))\Big], \tag{9}$$

where the $\omega_p^{\boldsymbol{M}}$ is also the positive-class weight for the alignment matrix $\boldsymbol{M}$. $\omega_p^{\boldsymbol{M}} = \Big[\sum_{i,j}(1-\boldsymbol{M}_{i,j})\Big] / \sum_{i,j} \boldsymbol{M}_{i,j}$.

Combining Equation 6 and Equation 9, we can derive the overall loss function we want to minimize as,

$$\mathcal{L} = \eta^{\boldsymbol{A}^*} \mathcal{L}_{rc}^* + \eta^{\boldsymbol{M}} \mathcal{L}_{axis} + \frac{\lambda}{2}\|\theta\|^2,$$
$$\text{where } \eta^{\boldsymbol{A}^*} = \frac{(n+m)^2}{2\sum_{i,j}(1-\boldsymbol{A}^*_{i,j})},\ \eta^{\boldsymbol{M}} = \frac{k^2}{2\sum_{i,j}(1-\boldsymbol{M}_{i,j})}. \tag{10}$$

$\eta^{\boldsymbol{A}^*}$ and $\eta^{\boldsymbol{M}}$ are the normalization terms to balance the scale of connectivity-enhanced reconstruction loss and axis guidance loss. $\frac{\lambda}{2}\|\theta\|^2$ is the L2 normalization for parameters. The trainable parameters include the weight matrices in GCNs as well as the one-init gate $g_{(i,j)}$ in Equation 5.

IV. Experiments

In this section, we evaluate the performance of the proposed SGVGAE model at performing ideology detection tasks on 6 real-world Twitter datasets. We compare the SGVGAE model with 12 state-of-the-art baselines based on accuracy, F1 score, and purity. We also assess the computational efficiency of the baselines. Additionally, we evaluate the performance of the SGVGAE model under different scales of semantic guidance. An ablation study is conducted to validate the impact of proposed components. The code of SGVGAE is available at https://github.com/jinningli/SGVGAE.

A. Datasets and Environments

The data was collected from the Twitter platform using keyword filters for the coarse-grained topics of the Philippines, the United States, and the South China Sea. Based on this coarse-grained dataset, we further split it into seven datasets of fine-grained topics using the PieClass algorithm [25]. PieClass is a weakly supervised topic classification model that facilitates topic splitting by providing topic-specific keywords as seeds. The statistics of the dataset are presented in Table I. We manual annotate part of the messages for evaluation.

The training and inference are conducted on a GPU machine equipped with 4 Nvidia RTX A5000 GPUs, 512GB of memory, and a 64-core CPU. For the GPT-3.5 and GPT-4 models, we utilize the OpenAI API to query the response, using a batch size of 20 to expedite the process.

B. Evaluation Metrics

We assess the ideology detection task as a binary classification task, employing metrics such as accuracy, macro F1-score, and average purity score. The macro F1-score is the average of the F1-scores for each class. The average purity score is also computed as the average for each class. We repeat each experiment with different random seeds 10 times and report the mean and standard deviation for the metrics. A smaller standard deviation indicates the model's stability under different initializations. For deterministic algorithms like zero-shot LLM encoders, we do not report the standard deviation.

In our experiments, we primarily focus on evaluating the belief classification of messages for which we have manual annotations. We also consider estimating the ground-truth user beliefs based on the beliefs expressed in the majority of their messages. As a result of this aggregation or averaging effect when considering multiple messages, the scores of user belief estimation tends to be better than those of individuals. Therefore, these results are omitted due to space limitations.

For all models that require semantic labels for semi-supervision, such as the proposed SGVGAE model and the Semi-RoBERTa model, we select the top 5% of the most popular tweets for semantic labeling, as introduced in Section III-C. We ensure those tweets with manual annotations reserved for evaluation and will not be selected for semi-supervision.

C. Baselines

- **VGAE-KM**: Variational graph auto encoders (VGAE) for graph embedding learning [9]. We apply KMeans (KM) to group the graph embeddings for ideology detection. The learning rate is set to 0.2 to adapt to our datasets, while other hyperparameters remain the same.
- **InfoVGAE**: Unsupervised Non-Negative VGAE [4] that maps graph nodes into an orthogonal embedding space. The ideology label is predicted by taking the maximum

TABLE II: Evaluation results for message ideology detection, in terms of accuracy, macro F1 score, and category-averaged purity metrics. The experiments were repeated 10 times, and we report the mean, and standard deviation for each metric.

Dataset	Crime			EDCA			Energy China		
Model Name	Acc. (%)	Macro F1 (%)	Purity (%)	Acc. (%)	Macro F1 (%)	Purity (%)	Acc. (%)	Macro F1 (%)	Purity (%)
VGAE-KM	$78.26^{\pm 6.14}$	$75.27^{\pm 6.76}$	$74.23^{\pm 6.44}$	$68.18^{\pm 6.46}$	$75.75^{\pm 9.04}$	$73.21^{\pm 10.0}$	$67.60^{\pm 14.4}$	$62.41^{\pm 17.8}$	$63.91^{\pm 16.6}$
InfoVGAE	$80.43^{\pm 7.04}$	$76.41^{\pm 8.73}$	$75.89^{\pm 8.47}$	$81.93^{\pm 12.0}$	$80.36^{\pm 14.5}$	$79.58^{\pm 13.9}$	$84.00^{\pm 12.6}$	$83.76^{\pm 16.8}$	$84.61^{\pm 16.2}$
RoBERTa-KM	$60.87^{\pm N/A}$	$58.01^{\pm N/A}$	$59.05^{\pm N/A}$	$52.27^{\pm N/A}$	$49.09^{\pm N/A}$	$50.42^{\pm N/A}$	$56.00^{\pm N/A}$	$50.98^{\pm N/A}$	$51.10^{\pm N/A}$
T-RoBERTa-KM	$89.13^{\pm N/A}$	$86.27^{\pm N/A}$	$87.25^{\pm N/A}$	$72.73^{\pm N/A}$	$66.67^{\pm N/A}$	$65.63^{\pm N/A}$	$70.00^{\pm N/A}$	$68.81^{\pm N/A}$	$69.49^{\pm N/A}$
TwhinBERT-KM	$50.00^{\pm N/A}$	$49.98^{\pm N/A}$	$63.48^{\pm N/A}$	$59.09^{\pm N/A}$	$52.86^{\pm N/A}$	$52.87^{\pm N/A}$	$52.00^{\pm N/A}$	$50.00^{\pm N/A}$	$71.43^{\pm N/A}$
Semi-RoBERTa	$85.87^{\pm 4.61}$	$79.51^{\pm 7.16}$	$88.65^{\pm 6.74}$	$71.59^{\pm 8.04}$	$67.07^{\pm 7.98}$	$66.93^{\pm 8.16}$	$66.00^{\pm 2.83}$	$58.97^{\pm 4.00}$	$64.55^{\pm 4.07}$
Semi-T-RoBERTa	$88.04^{\pm 1.54}$	$84.30^{\pm 1.71}$	$87.32^{\pm 3.00}$	$79.55^{\pm 1.92}$	$73.51^{\pm 2.03}$	$76.18^{\pm 2.93}$	$66.00^{\pm 2.83}$	$62.59^{\pm 3.81}$	$62.79^{\pm 3.52}$
Semi-TwhinBERT	$82.61^{\pm 12.3}$	$71.67^{\pm 24.1}$	$80.86^{\pm 14.5}$	$72.73^{\pm 3.21}$	$50.62^{\pm 13.13}$	$72.73^{\pm 3.21}$	$64.00^{\pm 2.97}$	$39.02^{\pm 11.3}$	$64.00^{\pm 3.05}$
Semi-TIMME	$81.64^{\pm 4.18}$	$77.19^{\pm 3.58}$	$76.27^{\pm 4.57}$	$82.78^{\pm 7.18}$	$81.42^{\pm 6.96}$	$82.32^{\pm 6.15}$	$86.51^{\pm 6.68}$	$86.94^{\pm 7.76}$	$87.29^{\pm 6.34}$
GPT-3.5	$85.87^{\pm 1.85}$	$81.81^{\pm 2.43}$	$83.52^{\pm 2.47}$	$81.82^{\pm 4.29}$	$79.68^{\pm 4.35}$	$78.74^{\pm 4.06}$	$92.80^{\pm 5.27}$	$92.01^{\pm 5.71}$	$93.65^{\pm 5.80}$
GPT-4	$76.09^{\pm 2.29}$	$69.57^{\pm 3.19}$	$70.35^{\pm 3.05}$	$80.68^{\pm \mathbf{1.61}}$	$79.21^{\pm \mathbf{1.60}}$	$78.64^{\pm \mathbf{1.36}}$	$83.60^{\pm 4.40}$	$83.04^{\pm 4.45}$	$83.08^{\pm 4.34}$
Mixtral-8x7B	$77.83^{\pm \mathbf{1.37}}$	$72.91^{\pm \mathbf{1.56}}$	$72.73^{\pm \mathbf{1.60}}$	$91.24^{\pm 2.11}$	$90.09^{\pm 2.37}$	$88.72^{\pm 2.31}$	$81.60^{\pm \mathbf{2.07}}$	$80.68^{\pm \mathbf{2.13}}$	$80.76^{\pm \mathbf{2.43}}$
SGVGAE (Ours)	$\mathbf{89.61}^{\pm 5.42}$	$\mathbf{86.87}^{\pm 6.85}$	$\mathbf{87.90}^{\pm 5.35}$	$\mathbf{93.18}^{\pm 2.58}$	$\mathbf{91.62}^{\pm 2.79}$	$\mathbf{92.71}^{\pm 3.51}$	$\mathbf{96.00}^{\pm 5.32}$	$\mathbf{95.54}^{\pm 6.38}$	$\mathbf{97.06}^{\pm 6.29}$
Dataset	US Military Philippine			Labor and Migration China			Social and Economic Issues		
Model Name	Acc. (%)	Macro F1 (%)	Purity (%)	Acc. (%)	Macro F1 (%)	Purity (%)	Acc. (%)	Macro F1 (%)	Purity (%)
VGAE-KM	$62.22^{\pm 4.40}$	$55.04^{\pm 4.60}$	$57.12^{\pm 3.45}$	$72.50^{\pm 7.93}$	$72.31^{\pm 8.28}$	$72.71^{\pm 7.82}$	$92.19^{\pm 8.07}$	$89.63^{\pm 14.95}$	$94.27^{\pm 4.49}$
InfoVGAE	$74.07^{\pm 5.30}$	$67.69^{\pm 4.53}$	$67.06^{\pm 3.20}$	$74.38^{\pm 5.47}$	$74.21^{\pm 5.60}$	$74.92^{\pm 5.46}$	$94.27^{\pm 0.74}$	$93.75^{\pm 0.84}$	$94.92^{\pm 0.54}$
RoBERTa-KM	$70.37^{\pm N/A}$	$57.14^{\pm N/A}$	$56.79^{\pm N/A}$	$62.50^{\pm N/A}$	$60.00^{\pm N/A}$	$66.67^{\pm N/A}$	$50.52^{\pm N/A}$	$49.87^{\pm N/A}$	$50.64^{\pm N/A}$
T-RoBERTa-KM	$51.85^{\pm N/A}$	$44.20^{\pm N/A}$	$48.33^{\pm N/A}$	$56.25^{\pm N/A}$	$51.52^{\pm N/A}$	$60.26^{\pm N/A}$	$86.60^{\pm N/A}$	$86.22^{\pm N/A}$	$85.81^{\pm N/A}$
TwhinBERT-KM	$51.85^{\pm N/A}$	$44.20^{\pm N/A}$	$48.33^{\pm N/A}$	$75.00^{\pm N/A}$	$74.60^{\pm N/A}$	$76.67^{\pm N/A}$	$73.20^{\pm N/A}$	$71.29^{\pm N/A}$	$71.57^{\pm N/A}$
Semi-RoBERTa	$72.22^{\pm \mathbf{2.62}}$	$65.60^{\pm \mathbf{1.50}}$	$66.36^{\pm 4.06}$	$53.13^{\pm 4.42}$	$43.86^{\pm 2.88}$	$63.33^{\pm 18.8}$	$86.08^{\pm 6.56}$	$85.75^{\pm 6.32}$	$86.09^{\pm 5.58}$
Semi-T-RoBERTa	$66.11^{\pm 4.45}$	$59.19^{\pm 5.08}$	$64.11^{\pm 9.51}$	$50.00^{\pm 3.56}$	$41.82^{\pm 4.03}$	$50.00^{\pm 4.20}$	$83.51^{\pm 4.37}$	$83.26^{\pm 4.32}$	$83.45^{\pm 3.73}$
Semi-TwhinBERT	$42.59^{\pm 18.3}$	$41.61^{\pm 17.4}$	$62.56^{\pm 3.03}$	$50.00^{\pm 13.4}$	$33.33^{\pm 19.2}$	$50.00^{\pm 4.98}$	$84.54^{\pm 1.46}$	$82.94^{\pm 2.26}$	$85.07^{\pm 0.13}$
Semi-TIMME	$73.18^{\pm 4.80}$	$66.38^{\pm 3.94}$	$65.37^{\pm 2.86}$	$74.83^{\pm 4.81}$	$75.09^{\pm 5.07}$	$76.22^{\pm 4.95}$	$95.74^{\pm 0.70}$	$95.23^{\pm 0.75}$	$95.42^{\pm 0.56}$
GPT-3.5	$55.93^{\pm 3.24}$	$50.28^{\pm 2.44}$	$54.56^{\pm \mathbf{1.23}}$	$75.63^{\pm 1.98}$	$74.18^{\pm 2.28}$	$\mathbf{82.97}^{\pm 2.41}$	$89.90^{\pm 1.87}$	$89.39^{\pm 1.98}$	$89.10^{\pm 1.94}$
GPT-4	$50.37^{\pm 5.30}$	$43.15^{\pm 3.70}$	$47.77^{\pm 2.03}$	$75.00^{\pm \mathbf{0.00}}$	$75.00^{\pm \mathbf{0.00}}$	$75.00^{\pm \mathbf{0.00}}$	$69.79^{\pm 5.14}$	$67.82^{\pm 5.89}$	$67.98^{\pm 5.72}$
Mixtral-8x7B	$55.55^{\pm 5.78}$	$49.99^{\pm 5.89}$	$54.40^{\pm 5.62}$	$69.38^{\pm 4.61}$	$68.40^{\pm 4.84}$	$71.96^{\pm 4.35}$	$81.13^{\pm 1.69}$	$80.51^{\pm 1.88}$	$80.19^{\pm 1.87}$
SGVGAE (Ours)	$\mathbf{77.78}^{\pm 4.46}$	$\mathbf{71.07}^{\pm 3.93}$	$\mathbf{69.44}^{\pm 2.66}$	$\mathbf{78.13}^{\pm 3.29}$	$\mathbf{78.01}^{\pm 3.34}$	$78.71^{\pm 3.26}$	$\mathbf{95.83}^{\pm 0.43}$	$\mathbf{95.50}^{\pm 0.49}$	$\mathbf{96.11}^{\pm 0.32}$

coordinate, eliminating the need for a clustering algorithm. The learning rate is set to 0.2 for consistency.

- **RoBERTa-KM**: We use the pre-trained RoBERTa model [23] (zero-shot) to encode tweet text. We apply the KMeans algorithm on the embedding of the *[CLS]* token in the last layer's hidden states to predict the ideology.
- **T-RoBERTa-KM**: We use the Twitter-RoBERTa model [11] pre-trained on a large-scale Twitter dataset to encode (zero-shot) tweet messages in our dataset. We take the *[CLS]* embedding from the last layer and apply KMeans.
- **TwhinBERT-KM**: We use the Twhin-BERT model [12], pre-trained on Twitter textual data and enhanced with a large-scale social graph. We take the *[CLS]* embedding and apply KMeans for ideology label prediction.
- **Semi-RoBERTa**: We fine-tune the RoBERTa model on our datasets with the same scale of GPT semantic labels used by our SGVGAE model for binary classification of belief ideology in the semi-supervised scenario.
- **Semi-T-RoBERTa**: We semi-supervisedly fine-tune the Twitter-RoBERTa model with the same scale of GPT semantic labels for belief classification.
- **Semi-TwhinBERT**: Similarly, we fine-tune the Twhin-BERT model with the same scale of GPT semantic labels.
- **Semi-TIMMIE**: A semi-supervised model [14] utilizing both graph structure and text features, with text features computed using BERT [22]. We apply the same scale of GPT belief labels as SGVGAE for semi-supervision.
- **GPT-3.5**: We use the OpenAI GPT-3.5 model [26] with a prompt for few-shot inference of the ideology prediction. The prompt is detailed in Section III-C.
- **GPT-4**: Similarly, we use the few-shot OpenAI GPT-4 model with the same prompt in Section III-C for the ideology prediction.
- **Mixtral-8x7B**: A pre-trained Mixture of Experts (MoE) model [27], Mixtral-8 $\times$ 7b, is adopted for few-shot inference of the ideology label. The prompt is the same as shown in Section III-C.

D. Evaluation Results

The evaluation results are presented in Table II. The proposed SGVGAE model outperforms the baselines across all six datasets. In terms of accuracy, macro F1 score, and average purity score of messages, SGVGAE achieves average improvements of 5.97%, 6.06%, and 6.51%, respectively, over the most competitive baseline. This demonstrates the effectiveness of the proposed semantic guidance mechanisms.

The performance of graph-based models such as VGAE-KM, InfoVGAE, TIMME, and SGVGAE is closely related to the *sparsity* of the social graph. For instance, they perform better on the EDCA and Social and Economic Issues datasets, which have average node degrees of 7.13 and 10.59, respectively. On these datasets, most graph-based models achieve comparable or better results than text-based models. However, on sparser datasets like the Crime, Labor, and Migration China datasets, with average node degrees of 5.81 and 3.18, VGAE-KM and InfoVGAE underperform compared to language models. In contrast, the SGVGAE model is more robust to sparse graphs. By leveraging semantic guidance from 5% of the messages, SGVGAE outperforms text-based algorithms, including the GPT-3.5 model used to compute the semantic labels for guidance. This highlights the importance of semantic guidance when the graph structure is sparse.

Compared to existing graph-based models such as VGAE-KM and InfoVGAE, the proposed SGVGAE demonstrates greater stability during training and under different random initializations of the model. For instance, the standard deviation of SGVGAE on the Energy China dataset is approximately $5\% - 7\%$, whereas the standard deviation of VGAE-KM and InfoVGAE is around $12\% - 17\%$. Typically, when the topology of the target social graph does not exhibit a clear separation, there may be many sub-optimal splits for the graph, leading to unstable results for graph-based models. In contrast, SGVGAE employs a small number of semantic labels to stabilize the training, resulting in consistently lower standard deviations compared to other graph-based models.

While benefiting from the semi-supervision of the belief label, the performance of the semi-supervised RoBERTa and the Twitter-RoBERTa model is still limited compared to SGVGAE. These two models focus solely on textual information, which is insufficient for ideology detection. The text of messages on social networks often contains randomness, as not all tweets express the users' beliefs or stances toward polarized topics. Additionally, these messages can be challenging for language-based models to understand and encode. In contrast, the graph structure provides a distilled or smoothed expression of users' opinions. The Semi-TwhinBERT is one of the models being enhanced with a social graph during pre-training, we observe improvements but it still yields sub-optimal performance because it is not designed to consider dataset-specific topology. The Semi-supervised TIMME model is the most competitive in the baselines, incorporating the social graph and utilizing additional belief labels as input for the node classification neural network. However, its performance is still limited due to the less effective node features encoded by BERT and the lack of interpretability. The proposed SGVGAE model uses a lightweight but effective supervision scheme to incorporate information from textual labels and maintains an interpretable representation space, resulting in consistent improvements across the datasets.

An additional advantage of the proposed semantic guidance techniques of SGVGAE is their compatibility with the interpretability of the belief representations. As a result, we do not require further clustering algorithms for the ideology prediction of nodes. Moreover, this property can be utilized for various downstream analysis tasks, such as assessing the strength of polarity for a tweet.

E. Computational Efficiency

We assess the computational efficiency of the models by reporting the number of parameters and the running time in Table III. Graph-centered models (VGAE-KM, InfoVGAE, Semi-TIMME, and SGVGAE) typically have an advantage over LLMs in terms of both training and inference time. Additionally, these models have a significantly smaller number of parameters, which requires less memory. This efficiency is an added benefit of graph-centered models.

TABLE III: Computational Efficiency

Model Name	# Trainable Param.	Time Cost
VGAE-KM	0.021M	4.17s
InfoVGAE	0.022M	4.03s
RoBERTa-KM	125M	12.25s
T-RoBERTa-KM	125M	12.21s
TwhinBERT-KM	279M	14.23s
Semi-RoBERTa	125M	55.21s
Semi-T-RoBERTa	125M	60.32s
Semi-TwhinBERT	279M	79.47s
Semi-TIMME	0.025M	9.57s
GPT-3.5	6B	22.69s
GPT-4	175B	99.45s
Mixtral-8x7B	46.7B	240.92s
SGVGAE (Ours)	0.022M	5.07s

F. Effect of Semantic Guidance Scale

In this subsection, we examine the impact of the scale of semantic labels on the performance of the SGVGAE model. We adjust the scale of semantic guidance from 0% to 20% and assess the performance of the SGVGAE model. The results are depicted in Figure 3. For most datasets, a larger scale of semantic guidance tends to improve performance, particularly for the purity metric. However, there are exceptions. For instance, in the case of the US Military Philippine dataset (shown in brown), increasing the scale of semantic-guidance labels leads to a significant decline in all metrics. This is due to the significant bias introduced by the GPT-3.5 model in this specific dataset (see Table II), which is used to generate belief labels for semantic guidance. Nevertheless, we still observe a performance improvement in this dataset when the guidance scale is small ($< 5\%$), which can be attributed to the gated fusion mechanism for connectivity guidance introduced in Section III-D. This mechanism allows the model to selectively integrate the semantic labels, mitigating the negative impact when the scale is limited.

G. Ablation Study

In this ablation study, we assess the impact of the connectivity guidance and axis guidance techniques of the SGVGAE model. We conduct experiments on all datasets, removing either or both of these techniques, and present the dataset-wise averaged metrics in Table IV. The removal of connectivity or axis guidance results in a decrease in metrics by approximately

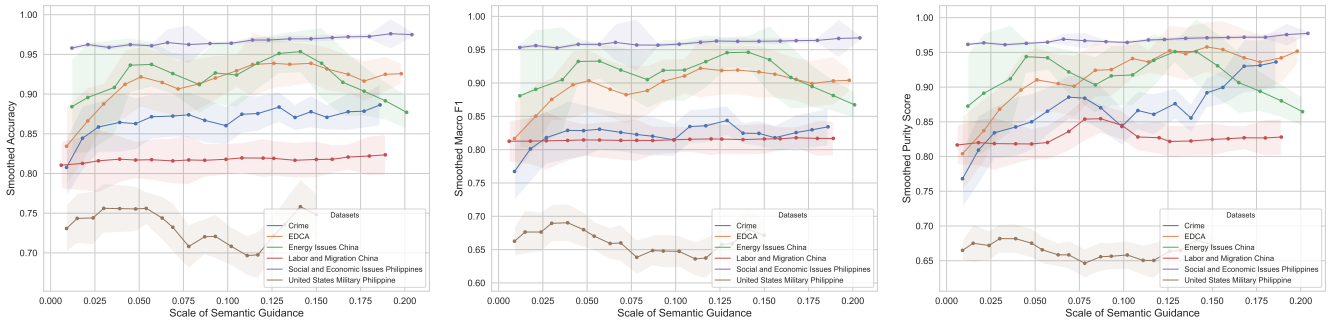

Fig. 3: The performance of the SGVGAE model under varying scales of semantic guidance for accuracy, macro F1, and purity, arranged from left to right. Each curve is smoothed to depict the overall trend as the scale of semantic guidance changes.

1.46% and 4.25%, respectively. This indicates that both connectivity and axis guidance are essential for the SGVGAE's performance. Furthermore, axis guidance appears to be more influential, as it directly optimizes node classification by aligning them with the axis representing the belief class. When both modules are removed, the SGVGAE model is reduced to a non-negative VGAE, leading to a substantial decrease.

TABLE IV: Result of Ablation Study. We report the average metrics across different datasets after removing the axis guidance, connectivity guidance, or both of them.

Model Name	Accuracy	Macro F1	Purity
No Guidance	82.73%	80.62%	80.70%
w/o Con. Guidance	87.29%	84.59%	85.58%
w/o Axis Guidance	84.44%	82.39%	82.24%
SGVGAE	88.42%	86.43%	86.98%

V. Related Works

A variety of representation learning techniques have been developed to encode social entities such as users and tweets. Two types of data are utilized to embed these entities into latent representations: the *structured social interaction graph*, and *unstructured multi-modal contexts* (e.g. texts and images).

The interaction history of users and messages, which encapsulates their beliefs and preferences, has been widely utilized to encode users and messages. Existing research often constructs interaction matrices and applies collaborative filtering [28], [29] or non-negative matrix factorization [6] to derive representations for users and messages. More recent studies have modeled the interaction history as graphs to derive social representations. The variational graph auto-encoders (VGAE) [9] introduce a framework with a GCN-based encoder and an inner-product decoder to map users and messages to a latent space with normal distribution variables. InfoVGAE [30] proposes a non-negative VGAE model capable of capturing the beliefs of users and messages in an interpretable social representation space. Subsequent work has also focused on multi-modal scenarios [31] and the disentanglement of representations [32]. However, existing interaction graph-based algorithms still face challenges related to sparsity, which can lead to instability or limited performance.

Recent advancements in large language models (LLMs) have significantly improved text-based social representation learning, as demonstrated by Twitter-RoBERTa [11] and TwhinBERT [12], particularly in zero-shot or few-shot scenarios. However, pure-textual models still face challenges in leveraging the structured knowledge of social graphs and concerns about computational efficiency due to the increasing complexity of LLMs. To address these issues, studies have investigated combining social graphs and message texts [13]: (i) Some research focuses on enhancing graph embeddings with LLM encoder features. In the TIMME [14] model, LLM features are used as input and semi-supervision labels support multi-task node classification. GraphBERT [15] uses textual embeddings of tweets, hashtags, and mentions as node features. LMKE [33] utilizes a contrastive learning framework to incorporate textual information in knowledge graph embedding (KGE) learning and enrich representation learning for long-tail entities. (ii) Other works focus on enhancing LLM encoding with graph node embeddings. KEPLER [16] jointly optimizes knowledge embedding and masked language modeling objectives, providing a unified model for text representations. NTULM [17] enriches the pre-trained BERT model by leveraging graph embeddings of multi-type non-textual units. In [34], the author proposes enhancing text-based belief response forecasting with graph embeddings.

However, embedding-based integration requires full textual information and may still lead to computational overhead. Therefore, in [21], the authors propose using LLM-generated labels to distantly supervise the training of graph models, which is fast and effective for the node classification task. Unlike node classification, this paper focuses on learning interpretable belief representations and developing an integration mechanism compatible with disentangled embeddings.

VI. Conclusion

This study presents a weakly-supervised semantic-guided graph representation learning model that integrates social graph and textual data for analyzing social beliefs in polarized networks. The model leverages the strengths of both graph and text, with optimization guided by soft textual labels generated by large language models. Key strategies include increasing connectivity among same-label nodes and enhancing align-

ment with specific belief axes, ensuring interpretability of the learned representations. Evaluation on six Twitter datasets demonstrates the model's effectiveness in stance detection, with notable improvements in accuracy, F1-score, and purity metrics. An ablation study validates the functionality of the proposed semantic guidance techniques. The possible future work includes integrating heuristic or active learning frameworks for the node selection strategy for LLM soft labeling.

ACKNOWLEDGEMENTS

Research reported in this paper was sponsored in part by the DARPA award HR001121C0165, the DARPA award HR00112290105, the DoD Basic Research Office award HQ00342110002, the Army Research Laboratory under Cooperative Agreement W911NF-17-20196. It was also supported in part by ACE, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

REFERENCES

[1] M. Li, X. Wang, K. Gao, and S. Zhang, "A survey on information diffusion in online social networks: Models and methods," *Information*, vol. 8, no. 4, p. 118, 2017.

[2] A. AlDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, vol. 58, no. 4, p. 102597, 2021.

[3] K. Darwish, P. Stefanov, M. Aupetit, and P. Nakov, "Unsupervised user stance detection on twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 141–152.

[4] J. Li, H. Shao, D. Sun, R. Wang, Y. Yan, J. Li, S. Liu, H. Tong, and T. Abdelzaher, "Unsupervised belief representation learning with information-theoretic variational graph auto-encoders," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1728–1738.

[5] C. Shang, R. Zhang, and X. Zhu, "The influence of social embedding on belief system and its application in online public opinion guidance," *Physica A: Statistical Mechanics and its Applications*, vol. 623, p. 128875, 2023.

[6] M. T. Al Amin, C. Aggarwal, S. Yao, T. Abdelzaher, and L. Kaplan, "Unveiling polarization in social networks: A matrix factorization approach," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[7] R. Dong, Y. Sun, L. Wang, Y. Gu, and Y. Zhong, "Weakly-guided user stance prediction via joint modeling of content and social interaction," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1249–1258.

[8] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[9] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *arXiv preprint arXiv:1611.07308*, 2016.

[10] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[11] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados, "Timelms: Diachronic language models from twitter," 2022.

[12] X. Zhang, Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, and A. El-Kishky, "Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter," *arXiv preprint arXiv:2209.07562*, 2022.

[13] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[14] Z. Xiao, W. Song, H. Xu, Z. Ren, and Y. Sun, "Timme: Twitter ideology-detection via multi-task multi-relational embedding," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 2258–2268.

[15] J. Wu, C. Zhang, Z. Liu, E. Zhang, S. Wilson, and C. Zhang, "Graphbert: Bridging graph and text for malicious behavior detection on social media," in *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2022, pp. 548–557.

[16] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, "Kepler: A unified model for knowledge embedding and pre-trained language representation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 176–194, 2021.

[17] J. Li, S. Mishra, A. El-Kishky, S. Mehta, and V. Kulkarni, "Ntulm: Enriching social media text representations with non-textual units," *arXiv preprint arXiv:2210.16586*, 2022.

[18] M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. S. Liang, and J. Leskovec, "Deep bidirectional language-knowledge graph pretraining," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 309–37 323, 2022.

[19] Z. Zhou and E. Elejalde, "Stance inference in twitter through graph convolutional collaborative filtering networks with minimal supervision," in *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 1030–1038.

[20] A. Danday and T. S. Murthy, "Twitter data analysis using distill bert and graph based convolution neural network during disaster," 2022.

[21] Z. Chen, H. Mao, H. Wen, H. Han, W. Jin, H. Zhang, H. Liu, and J. Tang, "Label-free node classification on graphs with large language models (llms)," *arXiv preprint arXiv:2310.04668*, 2023.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[24] M. Müller, M. Salathé, and P. E. Kummervold, "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter," *Frontiers in artificial intelligence*, vol. 6, p. 1023281, 2023.

[25] Y. Zhang, M. Jiang, Y. Meng, Y. Zhang, and J. Han, "Pieclass: Weakly-supervised text classification with prompting and noise-robust iterative ensemble training," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12 655–12 670.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[27] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[28] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Computer communications*, vol. 41, pp. 1–10, 2014.

[29] M. C. Pham, Y. Cao, R. Klamma, and M. Jarke, "A clustering approach for collaborative filtering recommendation using social network analysis." *J. Univers. Comput. Sci.*, vol. 17, no. 4, pp. 583–604, 2011.

[30] J. Li, H. Shao, D. Sun, R. Wang, Y. Yan, J. Li, S. Liu, H. Tong, and T. Abdelzaher, "Unsupervised belief representation learning with information-theoretic variational graph auto-encoders," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1728–1738. [Online]. Available: https://doi.org/10.1145/3477495.3532072

[31] L. Deng, Y. Huang, X. Liu, and H. Liu, "Graph2mda: a multi-modal variational graph embedding model for predicting microbe–drug associations," *Bioinformatics*, vol. 38, no. 4, pp. 1118–1125, 2022.

[32] J. Feng, L. Zhang, and L. Yang, "Concept-free causal disentanglement with variational graph auto-encoder," *arXiv preprint arXiv:2311.10638*, 2023.

[33] X. Wang, Q. He, J. Liang, and Y. Xiao, "Language models as knowledge embeddings," *arXiv preprint arXiv:2206.12617*, 2022.

[34] C. Sun, J. Li, Y. R. Fung, H. P. Chan, T. Abdelzaher, C. Zhai, and H. Ji, "Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting," *arXiv preprint arXiv:2310.13297*, 2023.