GPSocio: A Transformer-based General-purpose Social Network Representation System

Xinyi Liu[1] [0009-0003-4901-4796] ⋆, Dachun Sun[1] [0000-0003-4000-2783], and Tarek Abdelzaher[1] [0000-0003-3883-7220]

University of Illinois Urbana-Champaign, 201 N Goodwin Ave, Urbana, IL 61801, USA

Abstract. We present **GPSocio**[1], a general-purpose social network representation system designed to support diverse downstream analytics and enable knowledge transfer to data-scarce domains. While prior methods optimize embeddings for specific tasks, they often lack generalization. GPSocio leverages the emerging Graph Foundation Model (GFM) paradigm by aligning social graph structures with Large Language Models (LLMs) through post propagation sequences encoded in natural language. Extensive evaluations across four downstream tasks show that GPSocio consistently outperforms strong baselines, achieving **6.74%** gains in User macro-F1 for Sentiment Analysis, **21.29%** gains in User ARI for Ideology Classification, **14.75%** AUC improvement for Static Link Prediction, and **62.87%** User N@10 improvement for Temporal Link Prediction, demonstrating robust modeling of social semantics and diffusion dynamics.

Keywords: Social Network Representation Learning · Graph Foundation Models · Transformers for Social Networks · Domain-transferable Graph Embeddings, · Low-data Social Domain Transfer.

1 Introduction

Social network analysis underpins a wide range of tasks, such as user preference detection [29], information propagation prediction [41], ideology classification [24], and sentiment analysis [31]. Prior efforts largely focus on optimizing representations for specific tasks, often achieving strong in-domain performance but struggling to generalize beyond the training context. Such overfitting hampers generalization across evolving social contexts. Despite extensive efforts, no existing framework unifies semantic and diffusion information for cross-domain social reasoning.

The recent emergence of Graph Foundation Models (GFMs) offers new opportunities for learning transferable social representations through large-scale pretraining and domain adaptation. In particular, the integration of graph structures with Large Language Models (LLMs) enables richer semantic reasoning across nodes and edges [40]. However, a fundamental challenge remains: raw graph data must be reformulated into sequential, language-compatible formats to fully leverage LLM capabilities.

In this paper, we propose **GPSocio**, a transformer-based general-purpose framework for social network representation. GPSocio reformulates social graphs into *post*

⋆ This work was supported by DARPA (HR001121C0165, HR00112290105), DoD Basic Research Office (HQ00342110002), and ARL (W911NF-17-20196).

[1] Reproducible code is available at: https://github.com/tracy3057/GPSocio.

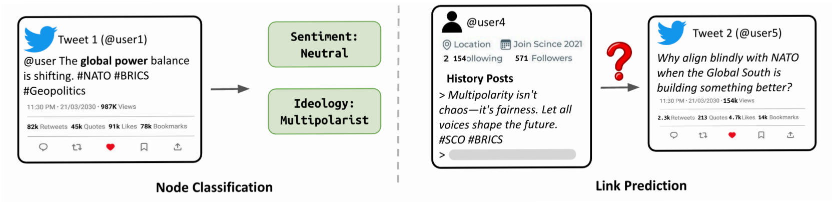
Tweet 1 (@user1)
@user The **global power** balance is shifting. #NATO #BRICS #Geopolitics
11:30 PM · 21/03/2030 · 987K Views
82k Retweets 45k Quotes 91k Likes 78k Bookmarks
Sentiment: Neutral
Ideology: Multipolarist
Node Classification
@user4
Location
Join Scince 2021
2 154ollowing
571 Followers
History Posts
> Multipolarity isn't chaos—it's fairness. Let all voices shape the future. #SCO #BRICS
>
?
Tweet 2 (@user5)
Why align blindly with NATO when the Global South is building something better?
11:30 PM · 21/03/2030 · 154k Views
2.3k Retweets 213 Quotes 4.7k Likes 14k Bookmarks
Link Prediction

Fig. 1: Illustration of GPSocio's downstream tasks: node classification (sentiment and ideology) and link prediction (static and temporal).

propagation sequences that capture both semantic content and diffusion dynamics in natural language form. It pre-trains representations via *contrastive next-user prediction* on large-scale sequences and refines them through *graph-aware domain adaptation* using target-specific interaction graphs. The overall architecture is illustrated in Figure 2.

Through this design, GPSocio bridges the gap between structured diffusion processes and language-based pretraining, enabling robust knowledge transfer across heterogeneous social domains. As social media ecosystems become increasingly fragmented and data-sparse, the need for unified social foundation models like GPSocio is becoming both urgent and inevitable.

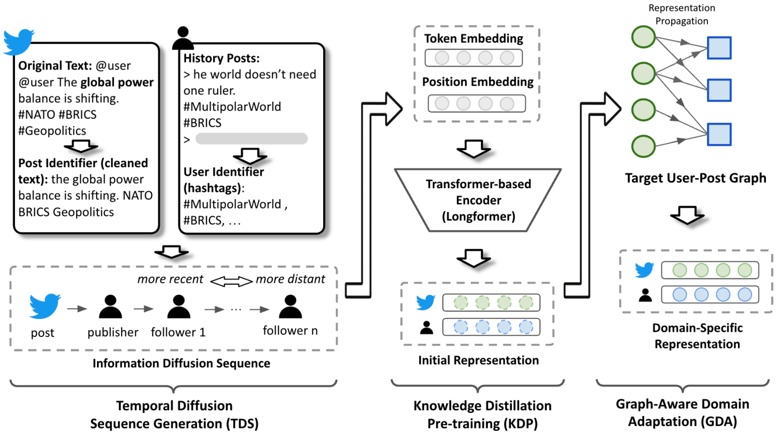
Original Text: @user @user The global power balance is shifting. #NATO #BRICS #Geopolitics
Post Identifier (cleaned text): the global power balance is shifting. NATO BRICS Geopolitics
History Posts:
> he world doesn't need one ruler. #MultipolarWorld #BRICS
>
User Identifier (hashtags): #MultipolarWorld , #BRICS, …
more recent
more distant
post
publisher
follower 1
follower n
Information Diffusion Sequence
Temporal Diffusion Sequence Generation (TDS)
Token Embedding
Position Embedding
Transformer-based Encoder (Longformer)
Initial Representation
Knowledge Distillation Pre-training (KDP)
Representation Propagation
Target User-Post Graph
Domain-Specific Representation
Graph-Aware Domain Adaptation (GDA)

Fig. 2: GPSocio framework overview, comprising: (1) temporal diffusion sequence generation (TDS), (2) knowledge distillation pre-training (KDP), and (3) graph-aware domain adaptation (GDA).

With the generated representations, GPSocio supports multiple downstream tasks (Figure 1), including node classification (Sentiment Analysis and Ideology Classification) and link prediction (Static and Temporal Link Prediction). Our main contributions are:

- We propose **GPSocio**, a general-purpose social network representation framework that unifies semantic and structural signals for diverse analytics tasks.
- We design a graph-to-language conversion using propagation sequences, hashtag profiles, and cleaned text to support LLM-based pretraining.
- We introduce a **next-user prediction** contrastive learning task that jointly models semantic affinity and diffusion behavior.

- We develop a lightweight **graph-aware domain adaptation** method to refine user embeddings with domain-specific structural information.
- GPSocio reduces inference time by **over 80%** compared to strong baselines like MINDS and RotDiff, while consistently delivering state-of-the-art performance.
- Extensive experiments demonstrate that GPSocio improves User macro-F1 by **6.74%** (SA) and User ARI by **21.29%** (Ideology Classification), and boosts AUC and User N@10 by **14.75%** and **62.87%** for link prediction, respectively.

2 Related Work

Social Network Representations Social networks support diverse downstream tasks, including user preference detection [29], information propagation prediction [41], ideology classification [24], sentiment analysis [31], and community detection [2]. Traditional approaches often rely on task-specific embeddings, such as key-value transformers for user history modeling [23] or GNNs for diffusion prediction [33, 36]. Other efforts adopt graph-based LSTMs for influencer detection [22] or attributed embeddings for community discovery [42]. LLM-based models have also been used in sentiment analysis [26], but most methods are tailored to single tasks or domains, limiting their generalization to heterogeneous social contexts.

Graph Foundation Models (GFMs) Inspired by the success of language and vision foundation models [5], recent work has explored Graph Foundation Models (GFMs) for cross-task and cross-domain transfer [25]. Existing GFMs can be categorized into three types:

GNN-based GFMs extend encoders like GCNs [20] with large-scale pretraining [17] or attention mechanisms [39], but are often limited to static structure modeling.

GNN+LLM hybrids combine structural and semantic features [8, 27], though the fusion is typically shallow and fails to enable deep language-informed reasoning. **LLM-based models** transform graphs into language-compatible sequences to leverage LLM capabilities [40], but often overlook temporal diffusion patterns crucial for modeling dynamic social networks.

Despite progress, existing GFMs rarely integrate semantic content and temporal diffusion in a unified, transferable representation. This remains a key gap that **GPSocio** seeks to address.

3 Problem Statement

We formally define the problem setting of GPSocio and the key concepts involved.

User (u): A unique account on a social media platform, identified by a user ID. We denote the source domain users as $\mathcal{U}^s = \{u_1^s, u_2^s, \dots, u_{N^s}^s\}$ and the target domain users as $\mathcal{U}^t = \{u_1^t, u_2^t, \dots, u_{N^t}^t\}$.

Post (p): A timestamped user-generated content item (e.g., a tweet) relevant to a domain of interest. The source posts are $\mathcal{P}^s = \{p_1^s, p_2^s, \dots, p_{M^s}^s\}$, and the target posts are $\mathcal{P}^t = \{p_1^t, p_2^t, \dots, p_{M^t}^t\}$.

Post Propagation Sequence (s): A temporal sequence capturing the diffusion of a post through users. For post p_i, its propagation sequence is $s_i = \{p_i, u_{i1}, u_{i2}, \dots\}$, where u_{i1} is the publisher followed by forwarders in reverse chronological order.

User History Sequence (h): A sequence of a user's historical post interactions, ordered reverse-chronologically as $h_i = \{u_i, p_{i1}, p_{i2}, \dots\}$.

Given the above definitions, the objective of **GPSocio** is to learn semantic- and diffusion-aware embeddings for:

- **Posts:** $r_{p_i^t}$ for each $p_i^t \in \mathcal{P}^t$
- **Users:** $r_{u_i^t}$ for each $u_i^t \in \mathcal{U}^t$

Our goal is to generate general-purpose representations that support diverse downstream tasks, e.g., sentiment classification, ideology classification, and link prediction, with minimal target-domain supervision. These embeddings should capture both semantic content and diffusion dynamics to enable robust generalization across heterogeneous, evolving social environments.

4 GPSocio Framework

We design three modules for the transformer-based GPSocio framework:

Temporal Diffusion Sequence Generation (TDS): To re-format social networks data into an LLM-understandable format, we adopt semantic identification for users and posts, and concat the items with certain chronological order as they propagate. This design captures the semantic information of users and posts with condensed text while retaining the information propagation patterns and hidden social relations.

Knowledge Distillation Pretraining (KDP): Given long information cascades, we adopt a modified Longformer [3] to handle extended sequences. GPSocio is pre-trained via *next user prediction* along post propagation sequences, which encode semantics, social ties, and behavioral cues—key to modeling engagement.

Graph-Aware Domain Adaptation (GDA) While social networks exhibit shared patterns, domain-specific adaptation refines representations to better capture target-domain nuances. We align the adaptation strategy with the downstream task.

4.1 Temporal Diffusion Sequence Generation (TDS)

The Temporal Diffusion Sequence Generation module extracts generalizable knowledge from data-rich source networks. RNNs [34] struggle with the long sequences in complex social settings, while GNNs [13] capture structure but often miss semantic content. To integrate both, we adopt a Transformer-based architecture, using a modified Longformer [3] as the backbone of **GPSocio**.

To extract knowledge for diffusion prediction and convert social data into a Longformer-compatible format, we define user and post identifiers as follows:

User Identifier: Each user is represented by the set of hashtags they have used, serving as compact, high-density markers of interests, ideology, and community ties. This concise representation shortens sequences while preserving identity cues.

Post Identifier: Each post is represented by its cleaned text, with URLs, mentions, and hashtags removed, capturing core semantics while reducing surface-level noise.

We then construct the **Post Propagation Sequence** s_i for each post p_i, which consists of:

$$X = \{[CLS], p_i, u_{i1}, u_{i2}, \dots\}. \tag{1}$$

Each sequence begins with the post p_i, followed by the publisher u_{i1} and subsequent forwarders ordered in **reverse-chronological order**.

This design offers several advantages:

- Reduces input sequence length and improves training efficiency by compactly representing users and posts.

- Preserves temporal diffusion dynamics through relative ordering, avoiding noise introduced by absolute timestamps.
- Emphasizes recent forwarders, who tend to show stronger engagement and influence on subsequent propagation, better modeling real-world information flow momentum.

By structuring the sequence to mirror the natural unfolding of social attention—from the post origin to its most recent amplifiers—we create a representation that improves the model's ability to generalize propagation behaviors across domains.

4.2 Knowledge Distillation Pretraining (KDP)

GPSocio adopts Longformer as the backbone model as it introduces a linear-scaling attention mechanism, making it well-suited for processing extended sequences while preserving global and local contextual dependencies.

Embedding Generation

To jointly capture semantic content and propagation dynamics, we construct token representations that combine two complementary components: token embeddings and position embeddings. This design integrates language model semantics [10] with temporal modeling from self-attention architectures [35], enabling the model to reason over both content and sequence order in diffusion cascades.

Each token, which represents either a user or a post, is mapped to a dense vector via a learned token embedding matrix $A \in \mathbb{R}^{D_w \times d}$, where D_w is the token vocabulary size and d is the embedding dimension. This captures core semantic features derived from textual identity and interaction history.

To preserve the temporal structure of cascades, we add position embeddings $B \in \mathbb{R}^{n \times d}$, where n denotes the maximum sequence length. Each position k is assigned an embedding $B_k \in \mathbb{R}^d$ that encodes its relative order, allowing the model to distinguish recent and early nodes—crucial for modeling diffusion momentum and engagement patterns.

The final input embedding for each token is given by:

$$E_w = \text{LayerNorm}(A_w + B_w), \tag{2}$$

where LayerNorm stabilizes the training process and promotes faster convergence [1].

GPSocio leverages Longformer [3] as its core encoder, which scales linearly with sequence length by combining local sliding attention and global attention to the [CLS] token. This architecture effectively captures both fine-grained diffusion signals and global semantic context in long propagation sequences.

Given a propagation sequence $s_i = p_i, u_{i1}, u_{i2}, \ldots$, we construct input embeddings $E_{s_i X} \in \mathbb{R}^{(h+1) \times d}$ by summing token and position embeddings (as defined previously), and pass them into the Longformer:

$$\begin{aligned} E_{s_i X} &= [E_{s_i[\text{CLS}]}, E_{s_i w_1}, \ldots, E_{s_i w_h}], \\ [r_{s_i[\text{CLS}]}, r_{s_i w_1}, \ldots, r_{s_i w_h}] &= \text{Longformer}(E_{s_i X}), \end{aligned} \tag{3}$$

where w_i are tokens (users or posts), and the [CLS] vector summarizes the sequence globally.

Post Representation: For each post p_i, we use the corresponding propagation sequence's [CLS] embedding to summarize its global diffusion dynamics:

$$r_{p_i} = r_{s_i[\text{CLS}]}. \tag{4}$$

User Representation: To construct each user's representation, we first apply Longformer to the user's personal post history s_{u_k}, yielding an initial semantic embedding $r_{u_{k_{\text{init}}}}$:

$$\begin{aligned} X &= [\text{CLS}], s_{u_k}, \\ r_{u_{k_{\text{init}}}} &= \text{Longformer}(E_{s_{u_k}} X). \end{aligned} \tag{5}$$

We then enrich this representation using the user's historical post embeddings, with more recent posts weighted quadratically to reflect their higher influence on current ideological stance [15, 14]:

$$r_{u_k} = \lambda r_{u_{k_{\text{init}}}} + \frac{1}{q} \sum_{i=1}^{q} i^2 r_{p_{ki}}, \tag{6}$$

where q is the number of historical posts, $r_{p_{ki}}$ is the embedding of the i-th post, and λ balances the static user profile and dynamic historical context.

This two-stage design enables GPSocio to capture both a user's general identity and recent ideological shifts, facilitating accurate modeling of evolving user behavior in dynamic social environments.

Model Pre-train with Longformer

We formulate model pre-training as a next-user prediction task within the post propagation sequence, effectively capturing propagation dynamics while maintaining a manageable sequence length. To optimize this task, we employ Item-Item Contrastive (IIC) learning [6]. Instead of random negative sampling, we enhance contrastive learning efficiency by leveraging in-batch next-item negatives [7], where negatives are drawn from ground-truth sequences within the same batch. This approach significantly reduces computational overhead while maintaining a low false-negative rate. The key insight behind this design is that social networks are inherently large and sparse—when users are arranged sequentially, the probability of two appearing in the same position due to forwarding or publishing the same post (i.e., a false negative) is extremely low.

The IIC Loss is defined as Equation 7 (τ is a constant):

$$\begin{aligned} sim_{p_i,u_k}(r_{p_i}, r_{u_k}) &= \frac{r_{p_i}^T r_{u_k}}{\|r_{p_i}\| \|r_{u_k}\|}, \\ L_{IIC} &= -log \frac{\frac{e^{sim(r_{p_i}, r_{true})}}{\tau}}{\Sigma_{u_k \in G_{batch}} e^{\frac{sim(r_{p_i}, r_{u_k})}{\tau}}}. \end{aligned} \tag{7}$$

4.3 Graph-Aware Domain Adaptation (GDA)

After pre-training, **GPSocio** enters fine-tuning to adapt its learned knowledge to structural and semantic nuances of the target domain. Specifically, we use a user-post bipartite graph $\mathcal{G}^t = (\mathcal{U}^t, \mathcal{P}^t, \mathcal{E}^t)$, where each edge $(u_k, p_i) \in \mathcal{E}^t$ indicates that user u_k interacted with post p_i (e.g., posted, forwarded, or liked).

To make user representations sensitive to domain-specific propagation, we refine them via graph-aware embedding propagation. Users often reflect the content they engage with, aggregating signals from associated posts enriches embeddings with contextual and topical cues relevant to the target network. Importantly, we only update user embeddings during this adaptation stage, while keeping post embeddings fixed.

This design choice stems from the observation that user ideologies and topical interests are dynamic, evolving over time based on external stimuli and personal expression patterns, whereas the semantic content of posts is static, anchored in their originally published text. Updating posts could introduce noise or distort their intrinsic semantic meaning, while refining users enables dynamic adaptation without compromising content integrity.

The user representation update rule is defined as:

$$r_{u_k} \leftarrow r_{u_k} + \alpha \sum_{p_i \in G^{u_k}_{\text{connected}}} r_{p_i}, \tag{8}$$

where r_{u_k} is the user embedding, r_{p_i} is the post embedding, $G^{u_k}_{\text{connected}}$ denotes the set of posts linked to user u_k, and α is a tunable propagation coefficient controlling the strength of information transfer from posts to users.

This domain-aware adaptation step enables the model to refine user representations by anchoring them to content-specific patterns in the target network, thereby reducing the semantic and structural shift between pre-trained representations and real-world diffusion behaviors. Ultimately, this improves the model's generalization and predictive capacity under low-data regimes.

5 Evaluation

We evaluate GPSocio across two categories of tasks (Figure 1). First, for node classification, we assess performance on Sentiment Analysis (SA) and Ideology Classification, where the goal is to predict node-level semantic or ideological attributes. Second, for link prediction, we evaluate both Static Link Prediction, which predicts the existence of links in the static graph, and Temporal Link Prediction, which predicts future interactions (i.e., next-item prediction) based on temporal graph dynamics.

5.1 Datasets

Table 1: Statistics of Pre-training and Target Datasets.

Dataset	Time Period	# Users	# Posts	Avg. Seq. Len	# Test Posts	# Comm. Users	Topic Sim.
Russia–Ukraine War	01 May 22–15 April 23	62,901	37,754	45.81	–	–	–
Attack Zelensky	01 May 22–15 April 23	12,171	1,668	32.67	731	3,182	0.286
Brics Superiority	01 May 22–15 April 23	1,845	481	18.36	124	1,183	0.309
Ukraine Nazi Claims	01 May 22–15 April 23	3,181	633	12.67	221	1,684	0.218

GPSocio was **pre-trained** on the *Russia-Ukraine War* dataset (keywords: "Russia Ukraine Conflict", etc.) and evaluated on three test sets: *Attack Zelensky* ("KievRegime"), *BRICS Superiority* ("BRICS"), and *Ukraine Nazi Claims* ("Azov Nazi"). Data was collected from Twitter (May 1, 2022 – April 15, 2023). We retained users with over 5 posts and posts propagated by at least 10 users. To assess domain shift, we computed cosine similarity between keyword sets. Dataset statistics appear in Table 1.

5.2 Metrics

For node classifications, we evaluate the performance over Macro-F1, Micro-F1 [30], Adjusted Rand Index (ARI) [18], and Normalized Mutual Information (NMI) [37].

- **Macro-F1** [30]. Calculates the F1 score for each class independently and averages them, treating all classes equally.

- **Micro-F1** [30]. Takes the F1 scores of each class and averages them, treating all classes equally.
- **ARI** [18]. Measures the similarity between two classifications by counting pairwise agreements, adjusting for random chance.
- **NMI** [37]. Quantifies the amount of shared information between predicted and true classifications, normalized by their entropies.

For link predictions, we evaluate the performance over Area Under the ROC Curve (AUC) [11], Average Precision (AP) [9], HIT@K and NDCG@K [12]. In this paper, the performance is evaluated at $K = 10, 20$, we use H@K and N@K as a short representation of the metrics.

- **AUC** [11]. Measures the ability of a model to distinguish between classes, plotting the true positive rate against the false positive rate at various thresholds.
- **AP** [9]. Computes the area under the precision-recall curve, summarizing the precision-recall trade-off across thresholds.
- **HIT@K (H@K)** [12]. Whether any of the top-K recommended items were in the test set for a given user.
- **NDCG@K (N@K)** [12]. NDCG is a widely used metric in information retrieval. It is used to calculate a cumulative score of an ordered set of items.

5.3 Baselines

We compare GPSocio against seven strong baselines spanning graph-based, sequence-based, and hybrid social representation models. To ensure fair comparison, all methods are evaluated under the same training/validation/test splits, and their embedding dimensions are uniformly set to 768.

- **Node2Vec** [16]: Learns node embeddings by simulating biased random walks and optimizing a neighborhood-preserving objective.
- **NDM** [38]: A neural diffusion model that combines attention and convolutional layers to model cascade dynamics under relaxed assumptions.
- **VGAE** [21]: A variational autoencoder with graph convolutions and an inner product decoder to reconstruct network links.
- **Inf-VAE** [33]: A variational framework that models social ties and activity sequences via co-attention and generative encoding.
- **MS-HGAT** [36]: A memory-based hypergraph attention network capturing user dependencies from friendships and cascades.
- **MINDS** [19]: Improves diffusion prediction via sequential hypergraphs and adversarial learning for better generalization.
- **RotDiff** [32]: Models social diffusion using hyperbolic attention and rotation-based encoding of temporal patterns.

5.4 Experimental Setting

GPSocio was pre-trained for 20 epochs using a batch size of 8, a temperature of 0.05, and a sequence window size of 512, with embedding dimension set to 768. Each propagation sequence was split into a **training set** (all nodes except the last three), a **validation set** (second-to-last node), and a **test set** (last node). During fine-tuning, the GDA propagation weight α was set to 0.5. For fair comparison, baseline models also use a representation dimension of 768.

Table 2: Performance Comparison of GPSocio and Baselines on Sentiment and Ideology Classification.

Dataset	Method	Sentiment Analysis				Ideology Classification			
		User macro-f1	Post macro-f1	User micro-f1	Post micro-f1	User ARI	Post ARI	User NMI	Post NMI
Attack Zelensky	VGAE	0.3209	0.3312	0.6271	0.6345	0.1834	0.1756	0.2347	0.2401
	InfVAE	0.3427	0.3496	0.6434	0.6548	0.2438	0.2489	0.2848	0.2937
	node2vec	0.3753	0.3802	0.6936	0.7021	0.3223	0.3278	0.3546	0.3591
	NeuralDiffusion	0.3215	0.3321	0.6844	0.7004	0.2970	0.3097	0.3484	0.3508
	MS-HGAT	0.3207	0.3259	0.6531	0.6607	0.4022	0.3936	0.3797	0.3902
	MINDS	0.3169	0.3279	0.6651	0.6567	0.4177	0.4092	0.3837	0.3914
	rotdiff (SOTA)	0.3128	0.3325	0.6641	0.6702	0.3557	0.3606	0.3803	0.4024
	GPSocio	**0.4051**	**0.4267**	**0.7982**	**0.8154**	**0.5286**	**0.5528**	**0.6102**	**0.6263**
Brics Superiority	VGAE	0.3584	0.3574	0.6105	0.6275	0.2112	0.2227	0.2540	0.2641
	InfVAE	0.3641	0.3734	0.6653	0.6792	0.3157	0.3219	0.3396	0.3582
	node2vec	0.3819	0.3907	0.6839	0.7021	0.2682	0.2886	0.3533	0.3694
	NeuralDiffusion	0.3427	0.3627	0.6532	0.6603	0.3347	0.3516	0.3472	0.3269
	MS-HGAT	0.3501	0.3618	0.6467	0.6567	0.3129	0.3225	0.3517	0.3702
	MINDS	0.3664	0.3729	0.6327	0.6770	0.3291	0.3054	0.3534	0.3818
	rotdiff (SOTA)	0.3681	0.3679	0.6682	0.6824	0.4020	0.4113	0.4285	0.4429
	GPSocio	**0.3997**	**0.4203**	**0.7669**	**0.7894**	**0.5190**	**0.5436**	**0.5737**	**0.5929**
Ukraine Nazi Claims	VGAE	0.2921	0.3032	0.5873	0.6057	0.3184	0.3208	0.2810	0.2927
	InfVAE	0.3327	0.3411	0.6086	0.6046	0.3582	0.3712	0.4259	0.4403
	node2vec	0.3769	0.3951	0.6675	0.6619	0.3223	0.3404	0.3944	0.4253
	NeuralDiffusion	0.3614	0.3741	0.6279	0.6354	0.2955	0.3139	0.3824	0.4028
	MS-HGAT	0.3668	0.3892	0.6356	0.6560	0.3764	0.3966	0.3442	0.3671
	MINDS	0.3693	0.3921	0.6285	0.6420	0.3679	0.4071	0.3698	0.4007
	rotdiff (SOTA)	0.3561	0.3745	0.6524	0.6876	0.4459	0.4629	0.4768	0.4748
	GPSocio	**0.4057**	**0.4253**	**0.7219**	**0.7497**	**0.4874**	**0.5175**	**0.5275**	**0.5407**

5.5 Node Classification Tasks

Task Formulation We consider two node classification tasks. For **Sentiment Analysis**, each node is labeled as positive, negative, or neutral using the NLTK toolkit [4]. In a weakly supervised setting, 10% of the labeled nodes are used for training. We extract GPSocio representations (refined through domain-specific adaptation) as input features and train a multi-layer perceptron (MLP) classifier to predict sentiment labels. For **Ideology Classification**, ideological labels (left-leaning, right-leaning, or neutral) are generated by GPT-4 via controlled prompting [28]. We apply principal component analysis (PCA) to reduce each 768-dimensional GPSocio embedding to 32 dimensions, followed by k-means classification with the number of clusters set to 3.

Overall Results Extensive results (Table 2) show that GPSocio consistently outperforms strong baselines across both node classification tasks. It improves user macro-F1 by an average of 7.6% in sentiment analysis and user ARI by 24.9% in ideology classification. On *Attack Zelensky*, for example, it raises macro-F1 from 0.3753 to 0.4051 and ARI from 0.4177 to 0.5286, demonstrating its strength in capturing both semantic information and structural dynamics.

5.6 Link Prediction Tasks

Table 3: Benchmarking GPSocio Against State-of-the-Art Methods on Static and Temporal Link Prediction Tasks.

Dataset	**Method**	**Static Link**		**Temporal Link**						**Inference Time**
		AUC (%)	AP (%)	User N@10	User H@10	User H@20	Post N@10	Post H@10	Post H@20	
Attack Zelensky	VGAE	72.36%	69.68%	0.1366	0.0848	0.1384	0.0107	0.0245	0.0614	9m43s
	InfVAE	74.58%	69.72%	0.0351	0.0372	0.0539	0.0156	0.0477	0.1399	10m28s
	node2vec	69.57%	65.44%	0.0980	0.1426	0.1616	0.0291	0.0719	0.1472	9m34s
	NeuralDiffusion	75.68%	73.93%	0.0759	0.1059	0.1462	0.0284	0.0728	0.1277	23m12s
	MS-HGAT	66.25%	64.32%	0.0507	0.0538	0.0744	0.0708	0.0865	0.1306	30m57s
	MINDS	68.74%	65.53%	0.0491	0.0527	0.0829	0.0680	0.0874	0.1407	26m24s
	rotdiff (SOTA)	69.97%	68.85%	0.0402	0.0609	0.0881	0.0538	0.0967	0.1529	7m11s
	GPSocio	**89.63%**	**88.68%**	**0.2026**	**0.2500**	**0.2815**	**0.1088**	**0.2059**	**0.3208**	**3m03s**
Brics Superiority	VGAE	64.48%	58.40%	0.0581	0.0658	0.0767	0.0700	0.1364	0.2727	5m43s
	InfVAE	68.89%	65.23%	0.0473	0.0688	0.0854	0.1195	0.2102	0.2757	7m25s
	node2vec	65.76%	63.74%	0.0711	0.1028	0.1583	0.0627	0.1875	0.3594	5m27s
	NeuralDiffusion	69.97%	68.29%	0.0751	0.0909	0.1145	0.0662	0.1164	0.2358	11m11s
	MS-HGAT	68.57%	65.72%	0.0892	0.0794	0.0966	0.0981	0.1151	0.1762	12m34s
	MINDS	66.36%	67.41%	0.0789	0.0851	0.1059	0.0867	0.1194	0.1627	9m47s
	rotdiff (SOTA)	72.51%	72.21%	0.0756	0.1069	0.1471	0.0824	0.1216	0.2675	3m21s
	GPSocio	**80.29%**	**79.72%**	**0.1650**	**0.2466**	**0.3007**	**0.1640**	**0.3184**	**0.5018**	**1m42s**
Ukraine Nazi Claims	VGAE	70.08%	68.52%	0.0712	0.0737	0.0773	0.0322	0.0754	0.1969	7m38s
	InfVAE	70.53%	67.61%	0.0692	0.0703	0.0918	0.0582	0.1313	0.1925	8m06s
	node2vec	72.07%	69.15%	0.0983	0.1426	0.1635	0.0695	0.1304	0.2275	7m55s
	NeuralDiffusion	75.44%	68.04%	0.1090	0.1175	0.1384	0.0450	0.1066	0.2306	10m37s
	MS-HGAT	73.31%	74.25%	0.1111	0.1209	0.1644	0.1351	0.1594	0.2207	19m41s
	MINDS	74.91%	73.96%	0.1007	0.1169	0.1782	0.1336	0.1729	0.2473	20m17s
	rotdiff (SOTA)	75.68%	73.67%	0.1029	0.1546	0.2103	0.1316	0.1998	0.2723	5m34s
	GPSocio	**86.97%**	**84.42%**	**0.1811**	**0.2916**	**0.3532**	**0.2178**	**0.4059**	**0.4989**	**2m39s**

Task Formulation We define static link prediction as a binary classification task, where the goal is to determine whether a link exists between a given user u_k and post p_i. For this, we obtain their representations r_{u_k} and r_{p_i}, and compute the cosine similarity as the link score (Eq. 7). We evaluate performance using 200 positive (existing) and 200 negative (non-existing) user-post pairs randomly sampled from the graph. A higher similarity score indicates a higher probability of link existence. In contrast, temporal link prediction (i.e., next-item prediction) is framed as a ranking task. For each user, we rank all candidate posts by their similarity to the user's representation and select the highest-scoring post as the predicted next interaction.

Overall Results As shown in Table 3, GPSocio consistently outperforms all baselines across both static and temporal link prediction tasks. For static link prediction, it achieves the highest AUC scores across all datasets, with improvements of **+19.0%** on *Attack Zelensky*, **+10.7%** on BRICS Superiority, and **+14.9%** on *Ukraine Nazi Claims* over the strongest baseline.

Performance gains are even more striking on the temporal link prediction task. GPSocio improves User N@10 by **48.3%**, **85.0%**, and **63.0%** respectively across the three datasets, demonstrating its superior ability to capture user-level temporal dynamics.

Significant improvements are also observed in Post N@10 and H@10, indicating robust modeling of post-level diffusion patterns.

In addition to predictive accuracy, GPSocio delivers substantial efficiency gains, reducing inference time by **57–84%** compared to strong baselines like RotDiff and MINDS. For instance, on *Attack Zelensky*, GPSocio completes inference in **3m03s**, versus 7m11s for RotDiff and 26m24s for MINDS. Similar trends hold for *BRICS Superiority* (**1m42s** vs. 3m21s/9m47s) and *Ukraine Nazi Claims* (**2m39s** vs. 5m34s/20m17s).

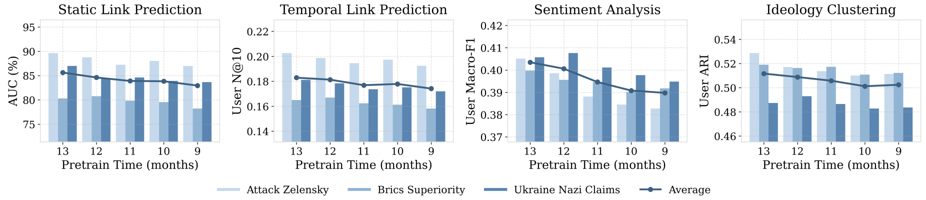

Fig. 3: Effect of Pre-training Duration on Performance Across Downstream Tasks.

5.7 Performance Analysis

GPSocio's strong results stem from three key innovations:

- Large-scale contrastive pretraining transfers social knowledge from high-resource to low-resource domains, addressing data scarcity.
- Diffusion-aware sequence modeling captures temporal dynamics and social semantics more effectively than graph-only or purely text-based approaches.
- Transformer-based encoding with Longformer integrates sequential and structural signals at scale, avoiding the computational overhead of traditional GNNs.

These innovations yield not only superior performance across diverse tasks but also high computational efficiency, making GPSocio a scalable and practical framework for real-world social analytics.

5.8 Robustness Study

Table 4: Robustness Study Across Temporal and User Overlap Variants.

Setting	Attack Zelensky				BRICS Superiority				Ukraine Nazi Claims			
	AUC	N@10	Macro-F1	ARI	AUC	N@10	Macro-F1	ARI	AUC	N@10	Macro-F1	ARI
Robustness Study: Varying Target Time Windows												
Full Time Window	**89.63%**	**0.2026**	**0.4051**	**0.5286**	**80.29%**	**0.1650**	0.3997	**0.5190**	86.97%	0.1811	0.4057	0.4874
Recent 4 Months	87.21%	0.1847	0.4004	0.5127	77.36%	0.1579	0.3862	0.5058	**87.11%**	0.1794	**0.4136**	0.4850
Recent 3.5 Months	88.96%	0.1807	0.4022	0.5203	79.43%	0.1604	**0.4018**	0.5098	85.84%	**0.1832**	0.3989	**0.4906**
Robustness Study: Effects of Pre-training Time Gaps with a 3-Month Target Time Window												
0-Day Gap	87.94%	0.2001	0.3966	0.5229	**82.42%**	**0.1748**	**0.4021**	**0.5042**	87.33%	0.2108	0.4145	0.5174
15-Day Gap	**88.02%**	0.2016	0.4027	**0.5418**	81.07%	0.1696	0.3846	0.4930	86.89%	0.2036	**0.4338**	**0.5257**
30-Day Gap	85.04%	**0.2128**	**0.4093**	0.5339	79.62%	0.1607	0.3947	0.4896	**87.51%**	**0.2243**	0.4142	0.5141
Robustness Study: Presence vs Absence of Common Users												
W/ Common Users	**89.63%**	0.2026	0.4051	0.5286	**80.29%**	0.1650	0.3997	0.5190	**86.97%**	0.1811	0.4057	0.4874
W/O Common Users	88.84%	**0.2074**	**0.4215**	**0.5437**	79.30%	**0.1685**	**0.4027**	**0.5269**	85.73%	**0.1903**	**0.4273**	**0.5046**

To assess GPSocio's resilience in real-world scenarios where data availability and alignment may vary, we evaluate its performance under reductions in pre-training data and perturbations in target domain conditions. We report four representative metrics across tasks: AUC (static link prediction), User N@10 (temporal link prediction), Macro-F1 (sentiment analysis), and ARI (ideology classification), chosen for their strong diagnostic value across task types.

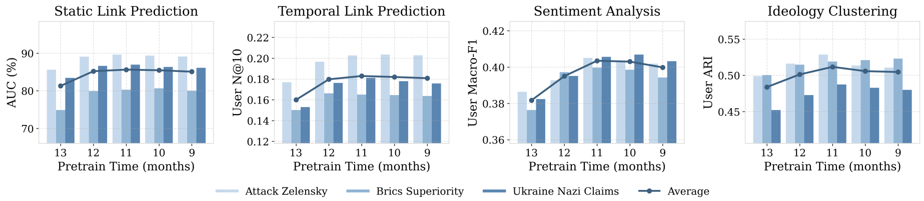

Fig. 4: Sensitivity of Performance Metrics to the Propagation Parameter α.

Pre-training Data Reduction. Shrinking the pre-training window from 13 to 9 months leads to negligible drops less than **2%** in AUC and N@10, and under **1.5%** in Macro-F1 and ARI (Figure 3). This demonstrates GPSocio's ability to retain generalization capacity even with significantly less pre-training data.

Target Domain Perturbations. We evaluate robustness under three common forms of distribution shift:

- **Shorter Target Windows:** Reducing the target window from the full time span to the most recent 3.5 months causes performance fluctuations of no more than **4.3%** across all metrics and datasets, demonstrating strong temporal stability (Table 4).
- **Increased Pretrain-Target Gap:** Introducing a gap of up to 30 days causes only minor variation, with peak performance at 15 days, indicating that GPSocio remains effective across time shifts.
- **Removal of Common Users:** Excluding shared users across domains does not harm performance. ARI on *Ukraine Nazi Claims*, for instance, improves from 0.4874 to 0.5046, indicating strong transferability beyond identity overlap.

Together, these results confirm that GPSocio is robust to data volume reductions and temporal or structural distribution shifts, owing to its combination of large-scale diffusion-aware pretraining and lightweight domain-specific refinement.

5.9 Sensitivity Study

To assess the stability of GPSocio under hyperparameter variations, we conduct a sensitivity analysis on the propagation weight α used during the domain-specific adaptation stage. This parameter controls the degree to which post embeddings influence user representations. A well-tuned α helps integrate content-specific signals without overwhelming the user's pre-trained semantics.

To streamline evaluation, we report four representative metrics: AUC (static link prediction), user N@10 (temporal link prediction), user Macro-F1 (sentiment analysis), and user ARI (ideology classification). These are chosen for their broad task coverage and strong correlation with other metrics within each task category, thus avoiding redundancy while maintaining diagnostic value.

As shown in Figure 4, GPSocio exhibits stable behavior across all tasks as α varies. Performance consistently improves as α increases from 0.1 to 0.5 or 0.7, indicating that moderate propagation effectively incorporates domain-specific context into user embeddings. Beyond 0.7, however, performance slightly declines, particularly in sentiment and temporal link prediction, suggesting that excessive propagation may introduce noise and dilute semantic distinctions.

Overall, GPSocio achieves optimal trade-offs between generality and adaptability when α is set between 0.5 and 0.7, demonstrating robust performance across downstream tasks without requiring fine-grained tuning.

Table 5: Ablation Results Comparing Performance With and Without GDA.

Dataset	W/ or W/O GDA	Static Link		Temporal Link		Sentiment Analysis		Ideology Clustering	
		AUC	AP	U. N@10	P. N@10	U. macro-f1	P. macro-f1	U. ARI	P. ARI
Attack Zelensky	W/O GDA	83.92%	83.72%	0.1533	0.0819	0.3572	0.3920	0.4753	0.5016
	W/ GDA	**89.63%** (+6.8%)	**88.68%** (5.9%)	**0.2026** (+32.2%)	**0.1088** (+32.8%)	**0.4051** (+13.4%)	**0.4267** (+8.9%)	**0.5286** (+11.2%)	**0.5528** (+10.2%)
Brics Superiority	W/O GDA	74.87%	73.14%	0.1501	0.1436	0.3763	0.4115	0.5004	0.5132
	W/ GDA	**80.29%** (+7.2%)	**79.72%** (9.0%)	**0.1650** (+9.9%)	**0.1640** (+14.2%)	**0.3997** (+6.2%)	**0.4203** (+2.1%)	**0.5190** (+3.7%)	**0.5436** (+5.9%)
Ukraine Nazi Claims	W/O GDA	83.43%	80.61%	0.1530	0.2004	0.3824	0.4142	0.4520	0.5039
	W/ GDA	**86.97%** (+4.2%)	**84.42%** (4.7%)	**0.1811** (+18.4%)	**0.2178** (+8.7%)	**0.4057** (+6.1%)	**0.4253** (+2.7%)	**0.4874** (+7.8%)	**0.5175** (+2.7%)

5.10 Ablation Study

To assess the impact of the Graph-Aware Domain Adaptation (GDA) module, we compare GPSocio's performance with and without GDA.

As shown in Table 5, even without GDA, GPSocio, trained solely on large-scale propagation sequences, already outperforms all baselines, confirming the strong generalization ability of its pre-trained embeddings.

Adding GDA further amplifies performance across all tasks. AUC improves by up to 6.06% (e.g., +6.8% on *Attack Zelensky*), and user N@10 gains reach as high as 20.2%, showing clear benefits in temporal link prediction. These results underscore that while pre-training captures transferable semantics and diffusion patterns, adapting to domain-specific graph structure is essential for optimal performance.

This validates our lightweight adaptation strategy (Section 4.3), which selectively updates user embeddings via post-to-user propagation—enhancing domain alignment without over-smoothing in sparse networks.

6 Conclusion

We propose GPSocio, a general-purpose framework that bridges semantic and diffusion signals via language-compatible propagation sequences and transformer modeling. Through contrastive pretraining and lightweight graph-aware adaptation, GPSocio enables robust, efficient transfer across diverse social tasks. Experiments on sentiment, ideology, and link prediction confirm its state-of-the-art performance and resilience under low-resource and shifting domains, underscoring the promise of language-informed social representation learning.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Bedi, P., Sharma, C.: Community detection in social networks. Wiley interdisciplinary reviews: Data mining and knowledge discovery **6**(3), 115–135 (2016)
3. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)

4. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc. (2009)
5. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
7. Chen, T., Sun, Y., Shi, Y., Hong, L.: On sampling strategies for neural network-based collaborative filtering. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 767–776 (2017)
8. Chien, E., Chang, W.C., Hsieh, C.J., Yu, H.F., Zhang, J., Milenkovic, O., Dhillon, I.S.: Node feature extraction by self-supervised multi-scale neighborhood prediction. arXiv preprint arXiv:2111.00064 (2021)
9. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Fawcett, T.: An introduction to roc analysis. Pattern recognition letters **27**(8), 861–874 (2006)
12. Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., Kashef, R.: Recommendation systems: Algorithms, challenges, metrics, and business opportunities. applied sciences **10**(21), 7748 (2020)
13. Feng, S., Zhao, K., Fang, L., Feng, K., Wei, W., Li, X., Shao, L.: H-diffu: hyperbolic representations for information diffusion prediction. IEEE Transactions on Knowledge and Data Engineering **35**(9), 8784–8798 (2022)
14. Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., Lorenz, J.: Models of social influence: Towards the next frontiers. Journal of Artificial Societies and Social Simulation **20**(4), 2 (2017)
15. Friedkin, N.E., Johnsen, E.C.: Social Influence Network Theory: A Sociological Examination of Small Group Dynamics. Cambridge University Press (2011)
16. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 855–864 (2016)
17. Huang, Q., Ren, H., Chen, P., Kržmanc, G., Zeng, D., Liang, P.S., Leskovec, J.: Prodigy: Enabling in-context learning over graphs. Advances in Neural Information Processing Systems **36** (2024)
18. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**, 193–218 (1985)
19. Jiao, P., Chen, H., Bao, Q., Zhang, W., Wu, H.: Enhancing multi-scale diffusion prediction via sequential hypergraphs and adversarial learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 8571–8581 (2024)
20. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
21. Kipf, T.N., Welling, M.: Variational graph auto-encoders. arXiv preprint arXiv:1611.07308 (2016)
22. Kumar, S., Mallik, A., Panda, B.: Influence maximization in social networks using transfer learning via graph-based lstm. Expert Systems with Applications **212**, 118770 (2023)
23. Li, J., Wang, M., Li, J., Fu, J., Shen, X., Shang, J., McAuley, J.: Text is all you need: Learning language representations for sequential recommendation. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1258–1267 (2023)

24. Li, J., Shao, H., Sun, D., Wang, R., Yan, Y., Li, J., Liu, S., Tong, H., Abdelzaher, T.: Unsupervised belief representation learning with information-theoretic variational graph autoencoders. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1728–1738 (2022)
25. Liu, J., Yang, C., Lu, Z., Chen, J., Li, Y., Zhang, M., Bai, T., Fang, Y., Sun, L., Yu, P.S., et al.: Towards graph foundation models: A survey and beyond. arXiv preprint arXiv:2310.11829 (2023)
26. Liu, W., Wen, B., Gao, S., Zheng, J., Zheng, Y.: A multi-label text classification model based on elmo and attention. In: MATEC Web of Conferences. vol. 309, p. 03015. EDP Sciences (2020)
27. Mavromatis, C., Ioannidis, V.N., Wang, S., Zheng, D., Adeshina, S., Ma, J., Zhao, H., Faloutsos, C., Karypis, G.: Train your own gnn teacher: Graph-aware distillation on textual graphs. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 157–173. Springer (2023)
28. OpenAI: Gpt-4 technical report. https://arxiv.org/abs/2303.08774 (2023)
29. Pereira, F.S., Gama, J., de Amo, S., Oliveira, G.M.: On analyzing user preference dynamics with temporal social networks. Machine Learning **107**, 1745–1773 (2018)
30. Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020)
31. Pozzi, F.A., Fersini, E., Messina, E., Liu, B.: Sentiment analysis in social networks. Morgan Kaufmann (2016)
32. Qiao, H., Feng, S., Li, X., Lin, H., Hu, H., Wei, W., Ye, Y.: Rotdiff: A hyperbolic rotation representation model for information diffusion prediction. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 2065–2074 (2023)
33. Sankar, A., Zhang, X., Krishnan, A., Han, J.: Inf-vae: A variational autoencoder framework to integrate homophily and influence in diffusion prediction. In: Proceedings of the 13th international conference on web search and data mining. pp. 510–518 (2020)
34. Sherstinsky, A.: Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena **404**, 132306 (2020)
35. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., Jiang, P.: Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In: Proceedings of the 28th ACM international conference on information and knowledge management. pp. 1441–1450 (2019)
36. Sun, L., Rao, Y., Zhang, X., Lan, Y., Yu, S.: Ms-hgat: memory-enhanced sequential hypergraph attention network for information diffusion prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 4156–4164 (2022)
37. Vinh, N., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants. Properties, Normalization and Correction for Chance **18** (2009)
38. Yang, C., Sun, M., Liu, H., Han, S., Liu, Z., Luan, H.: Neural diffusion model for microscopic cascade prediction. arXiv preprint arXiv:1812.08933 (2018)
39. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.Y.: Do transformers really perform badly for graph representation? Advances in neural information processing systems **34**, 28877–28888 (2021)
40. Zhao, H., Liu, S., Chang, M., Xu, H., Fu, J., Deng, Z., Kong, L., Liu, Q.: Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. Advances in Neural Information Processing Systems **36** (2024)
41. Zhao, J., Wu, J., Feng, X., Xiong, H., Xu, K.: Information propagation in online social networks: a tie-strength perspective. Knowledge and Information Systems **32**, 589–608 (2012)
42. Zhou, X., Su, L., Li, X., Zhao, Z., Li, C.: Community detection based on unsupervised attributed network embedding. Expert Systems with Applications **213**, 118937 (2023)