An LLM-Embedding Semantic Adaptation Network for Post-level Semantic Drift Evaluation

Anonymous ACL submission

Abstract

Evaluating semantic drift is essential for understanding dynamical discourse evolution and opinion formation in online discussions. However, sparse and uneven distributions of event-specific keywords prevent traditional models from capturing post-level semantic drift. Thus, to address this issue, we propose an LLM-embedding Semantic Adaptation Network (LLM-SAN), which is a hybrid semantic drift evaluation model with an LLM-Embedding gated recurrent unit (GRU) module, an LLM-Embedding graph convolutional network (GCN) module and a multi-expert adaptive fusion module. The GRU module is used to extract features from event related posts, and The GCN is used to extract features from temporal graphical topic posts. Then, the features are merged by the multi-expert adaptive fusion module. Finally, this module predicts the future post embedding, and the prediction error is used to evaluate and detect the semantic drift points. Extensive experiments are conducted, and the results show that LLM-SAN achieves the state-of-the-art performance on the semantic drift evaluation task, compared to the other baselines. Ablation experiments are also conducted to show the effectiveness of each module in LLM-SAN.

1 Introduction

Semantic drift evaluation is a critical yet challenging task to quantify and measure dynamical semantic changes, as illustrated in Fig. 1. In this task, event-related posts causally evolve and drift with time-varying distribution within public attention and discussion topics (Wu et al., 2024; Wang and Goutte, 2018; Balepur et al., 2023; Kang et al., 2024; He et al., 2024). The vocabulary distributions, sentiment polarity, topic structures and contextual semantic associations gradually change during the evolution (Garroppo et al., 2018; Karjus et al., 2020; Gaul and Vincent, 2017; Unankard and Nadee, 2020; Cao et al., 2023; Periti and Tahmasebi, 2024).

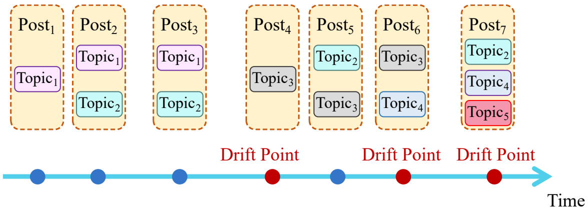

Figure 1: Semantic Drift Evauation.

There have been studies on event-related content evolution, which can be classified into three types: (i) probabilistic models (Blei et al., 2003; Blei and Lafferty, 2006; Wang and McCallum, 2006) learning topic changes from word-frequency distributions; (ii) incremental models (Wang and Goutte, 2018; Rieger et al., 2022; Patil et al., 2025b) learning dynamic semantic evolution via online detection mechanisms; and (iii) deep-learning models (Wang et al., 2019; Dieng et al., 2019; Rahimi et al., 2024) learning semantic representations through neural encoders.

However, these models all perform at the time level or event level. Thus, they hardly model the rapidly-evolving post content, where the semantic drift is reflected through sparse and unevenly distributed event-specific keywords. This makes them difficult to leverage fine-grained semantic information and perform with low accuracy if event scales vary significantly or salient semantic content is unevenly distributed.

Thus, to address this issue, we propose an LLM-Embedding Semantic Adaptation Network (LLM-SAN) for post-level semantic drift evaluation. In this model, temporal dynamics features are obtained from a gated recurrent unit (GRU) with inputs from an LLM-based encoder, capturing accumulated semantic drift across posts. Meanwhile, topical features are obtained from a graph convolutional network (GCN) with inputs from an

LLM-based generator, enhancing topic representations and sparse keyword signals. An attention-guided multi-expert mechanism then adaptively fuses these features, which are subsequently used for event-stage segmentation.

Thus, The main contributions of this work are as follows:

1. LLM-SAN model is proposed to evaluate the post-level semantic drift, providing a systematic way to capture fine-grained semantic changes in individual posts.
2. The model uses a GRU with inputs from an LLM-based encoder to capture accumulated post-level semantic drift, a GCN with inputs from an LLM-based generator to enhance topic representations and discontinuous keyword signals, and an attention-guided multi-expert fusion mechanism to adaptively combine multi-scale features, effectively addressing sparse or unevenly distributed event-specific keywords.
3. Experiments demonstrate that LLM-SAN achieves state-of-the-art performance on post-level semantic drift detection. Ablation studies verify the effectiveness of the key modules, and tests on real-world datasets confirm its validity, ultimately facilitating accurate event stage segmentation.

2 Related Work

This section introduces related work on semantic drift and content evolution, including probabilistic models, incremental models, and deep-learning models.

Probabilistic models describe the documents by using latent topics inferred from word-frequency distributions. Among them, latent dirichlet allocation (LDA) (Blei et al., 2003; Chauhan and Shah, 2021; Goyal and Kashyap, 2022) models document-topic and topic-word distributions under Dirichlet priors. Then, dynamic topic models (Blei and Lafferty, 2006) extend LDA by linking topic parameters across adjacent time slices through state-space assumptions. Further, topics over time (Wang and McCallum, 2006) incorporate temporal priors to capture time-dependent topic distribution changes.

Incremental models are designed to track the evolution of semantic in streaming or continuously arriving data. Among them, online LDA (Wang and Goutte, 2018; Fan et al., 2021; Zhou et al., 2023; Balepur et al., 2023) models are proposed to update topic parameters incrementally as new documents arrive sequentially. By contrast, Rolling LDA (Rieger et al., 2022) uses a sliding window to capture the features of continue rolling topics. Moreover, joint dynamic topic model (Zhang and Lauw, 2022) is also proposed to analyze varying topics by the combination of time-aware optimal transport and temporal point process techniques. Graph-based online models (Patil et al., 2025b,a) are proposed to represent topic drift through evolving graph structures.

Deep-learning models leverage neural encoders to learn semantic representations. Among them, neural topic models (Wang et al., 2019; Wu et al., 2024; Boutaleb et al., 2024) are proposed to perform topic inference by using neural variational framework. Based on this framework, dynamic embedded topic model (DETM) (Dieng et al., 2019) use word embedding technique to capture the features of topic evolution over time. Then, dynamic structured neural topic model (Miyamoto et al., 2023) is proposed to capture the topic branching and merging features in evolution process by self-attention mechanisms. Further, aligned neural topic model (ANTM) (Rahimi et al., 2024) uses large language model (LLM) to extract time-aware features from evolving topics.

3 Methodology

The overall architecture of the proposed LLM-SAN model is presented as shown in Fig. 2.

3.1 Temporal Evolution Trend Extraction

Formally, let an event consist of a sequence of temporally ordered posts $\mathcal{P}_{1:T} = \{p_1, p_2, \ldots, p_T\}$, where each post p_t is associated with textual content and time step t.

For individual post, an LLM-based encoder is applied to generate single post content semantic embeddings:

$$\mathbf{e}_t^{\text{temporal}} = \text{LLM}_{\text{emb}}(p_t), \tag{1}$$

$\mathbf{e}_t^{\text{temporal}} \in \mathbb{R}^{d_1}$ is the semantic embedding of p_t. The resulting embedding $\mathbf{e}_t^{\text{temporal}}$ is subsequently input into a GRU module to capture temporal evolution trends and accumulate post-level semantic drift.

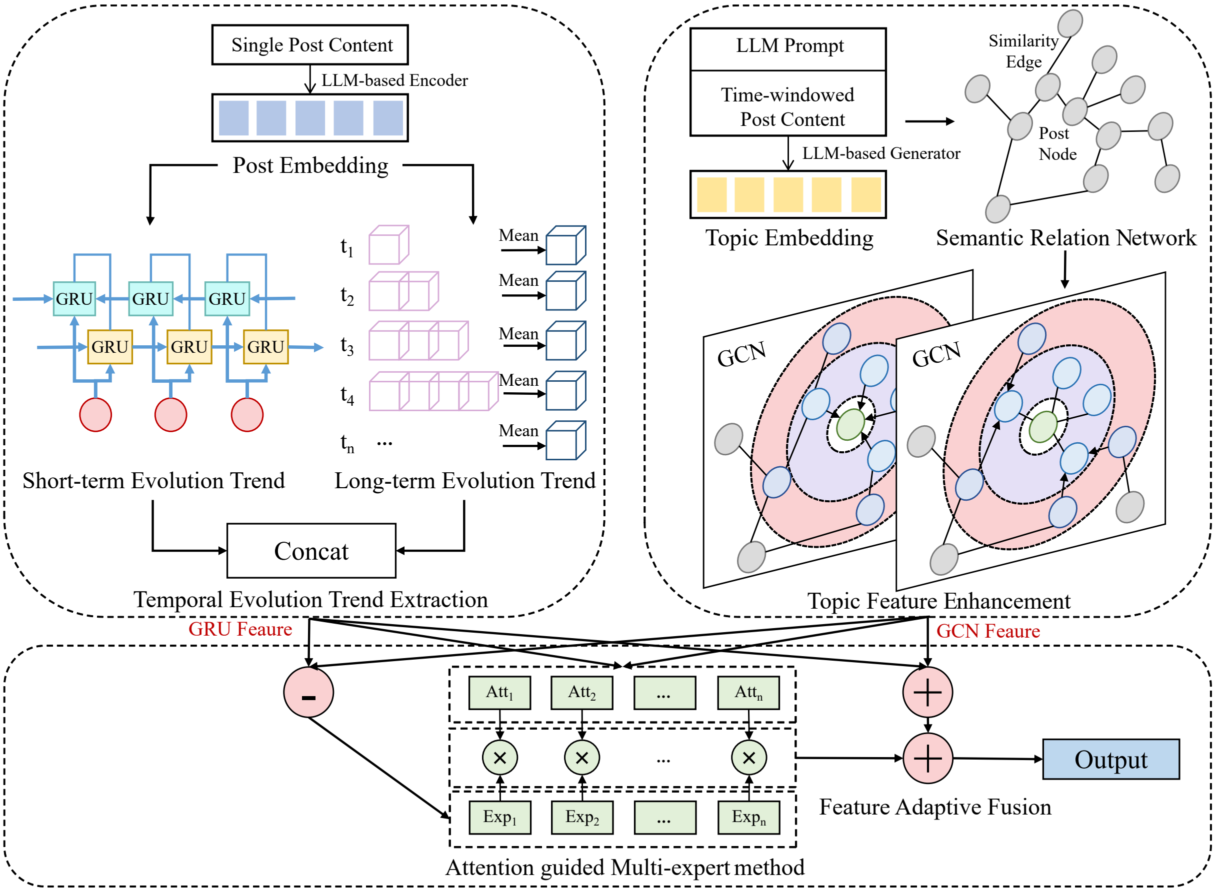

Figure 2: LLM-Embedding Semantic Adaptation Network.

At each time step t, a two-layer GRU processes the sequence of post embeddings, $\mathbf{e}_{t-c}^{\text{temporal}}, \ldots, \mathbf{e}_{t-1}^{\text{temporal}}$, from the preceding $c = 1, 2, \ldots, T-1$ time steps, updating its hidden state recurrently to capture short-term evolution trends in event-related posts.

Specifically, given the post embedding at time step $t-1$, the GRU hidden state is updated as

$$\mathbf{z}_t = \sigma\left(\mathbf{W}_z \mathbf{e}_{t-1}^{\text{temporal}} + \mathbf{U}_z \mathbf{h}_{t-1}\right), \tag{2}$$

$$\mathbf{r}_t = \sigma\left(\mathbf{W}_r \mathbf{e}_{t-1}^{\text{temporal}} + \mathbf{U}_r \mathbf{h}_{t-1}\right), \tag{3}$$

$$\tilde{\mathbf{h}}_t = \tanh\left(\mathbf{W}_h \mathbf{e}_{t-1}^{\text{temporal}} + \mathbf{U}_h \left(\mathbf{r}_t \odot \mathbf{h}_{t-1}\right)\right), \tag{4}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t, \tag{5}$$

where $\mathbf{h}_{t-1} \in \mathbb{R}^{d_h}$ is the GRU hidden state from the previous time step, with hidden dimension d_h, and the initial hidden state $\mathbf{h}_0$ is set to a zero vector. $\mathbf{W}_z \in \mathbb{R}^{d_h \times d_1}$, $\mathbf{W}_r \in \mathbb{R}^{d_h \times d_1}$, and $\mathbf{W}_h \in \mathbb{R}^{d_h \times d_1}$ are learnable weight matrices applied to the input embedding $\mathbf{e}_{t-1}^{\text{temporal}}$, where d_h denotes the hidden dimension and d_1 denotes the dimensionality of $\mathbf{e}_t^{\text{temporal}}$; $\mathbf{U}_z \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{U}_r \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{U}_h \in \mathbb{R}^{d_h \times d_h}$ are learnable weight matrices applied to the previous hidden state, with dimension d_h; $\mathbf{z}_t \in \mathbb{R}^{d_h}$ and $\mathbf{r}_t \in \mathbb{R}^{d_h}$ are the update and reset gates, controlling information flow; $\tilde{\mathbf{h}}_t \in \mathbb{R}^{d_h}$ is the candidate hidden state, and $\mathbf{h}_t \in \mathbb{R}^{d_h}$ is the updated hidden state; $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the sigmoid function, $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ denotes the hyperbolic tangent function, and $\odot$ represents element-wise multiplication; Through recursive updates over the previous c time steps (i.e., using embeddings from time steps $t-c$ to $t-1$), $\mathbf{h}_t$ captures short-term evolution trends in event-related posts.

To balance short-term variations with long-term evolution trends, a historical mean embedding is computed over all past posts up to time $t-1$ as

$$\bar{\mathbf{e}}_t = \frac{1}{t-1} \sum_{i=1}^{t-1} \mathbf{e}_i^{\text{temporal}}, \tag{6}$$

and mapped through a nonlinear transformation:

$$\mathbf{m}_t = \mathrm{ReLU}(\mathbf{W}_m \bar{\mathbf{e}}_t + \mathbf{b}_m), \tag{7}$$

where $\mathbf{W}_m \in \mathbb{R}^{d_h \times d_1}$ and $\mathbf{b}_m \in \mathbb{R}^{d_h}$ are learnable weight matrices applied to the historical mean embedding, and $\mathbf{m}_t \in \mathbb{R}^{d_h}$ captures the long-term semantic evolution trend accumulated before time step t.

Then, the short-term representation $\mathbf{h}_t$ and the long-term trend representation $\mathbf{m}_t$ are concatenated and projected via a linear layer to obtain the GRU output representation as follows:

$$\mathbf{f}_t^{\mathrm{GRU}} = \mathbf{W}_g[\mathbf{h}_t; \mathbf{m}_t] + \mathbf{b}_g, \tag{8}$$

where $[\,;\,]$ denotes vector concatenation, and $\mathbf{W}_g \in \mathbb{R}^{d_1 \times 2d_h}$ and $\mathbf{b}_g \in \mathbb{R}^{d_1}$ are learnable weight matrices applied to the concatenated embeddings. $\mathbf{f}_t^{\mathrm{GRU}} \in \mathbb{R}^{d_1}$ is a representation that jointly preserves long-term stable semantic information while emphasizing short-term semantic shifts, thereby providing a comprehensive characterization of post-level semantic drift.

3.2 Topic Feature Enhancement

To encode contextual topic semantics from a preceding time window, a LLM-based generator is applied to the time-windowed post contents $\mathcal{P}_{t-c:t-1} = \{p_{t-c}, p_{t-c+1}, \dots, p_{t-1}\}$, where $c = 1, 2, \dots, T-1$ denotes the length of the temporal window. The LLM-based generator produces a topic-aware embedding:

$$\mathbf{e}_t^{\mathrm{topic}} = \mathrm{LLM}_{\mathrm{gen}}(\mathcal{P}_{t-c:t-1}). \tag{9}$$

where $\mathbf{e}_t^{\mathrm{topic}} \in \mathbb{R}^{d_2}$, here d_2 is the dimension of the mean-pooled output of the generator LLM over the window.

The generation process is guided by the following prompt:

"You are a Twitter user. Based on the following posts published in the previous time period, generate a plausible post for the next time step."

A single vector summarizing the window is obtained via mean pooling over the hidden states. The resulting embedding $\mathbf{e}_t^{\mathrm{topic}}$ $, t = c+1, c+2, \dots, T$ is then input into a GCN module to enhance contextual topic representations.

After obtaining the topic-aware embeddings $\mathbf{e}_t^{\mathrm{topic}}$ from the preceding time window, these embeddings are treated as nodes in a temporal topic graph, and a semantic relation network is constructed to enhance discontinuous keyword signals and topic-level representations using a two-layer GCN.

First, a pairwise similarity matrix $\mathbf{S}$ is computed among the topic embeddings:

$$\mathbf{S} = \mathbf{E}^{\mathrm{topic}} \left(\mathbf{E}^{\mathrm{topic}}\right)^{\top}, \tag{10}$$

where $\mathbf{E}^{\mathrm{topic}} = [\mathbf{e}_{c+1}^{\mathrm{topic}}; \mathbf{e}_{c+2}^{\mathrm{topic}}; \dots; \mathbf{e}_T^{\mathrm{topic}}] \in \mathbb{R}^{(T-c) \times d_2}$ denotes the matrix of topic embeddings corresponding to historical time steps within the preceding window. Edges are formed between node pairs corresponding to the top 5% largest values in the similarity matrix $\mathbf{S}$, resulting in an undirected semantic graph $\mathcal{G}$. Each node represents a timespecific topic embedding $\mathbf{e}_t^{\mathrm{topic}}$, and edges denote semantic similarity between topics across time steps.

Subsequently, a two-layer GCN is applied to aggregate information from semantically related nodes, yielding enhanced topic representations. The update rule at layer $l+1$ is defined as:

$$\mathbf{H}^{(l+1)} = \sigma\Big(\mathbf{D}^{-\frac{1}{2}} \tilde{\mathbf{A}} \mathbf{D}^{-\frac{1}{2}} \mathbf{H}^{(l)} \mathbf{W}_p^{(l)}\Big), \tag{11}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ denotes the adjacency matrix $\mathbf{A} \in \mathbb{R}^{(T-c)\times(T-c)}$ with self-loops $\mathbf{I} \in \mathbb{R}^{(T-c)\times(T-c)}$, $\mathbf{D} \in \mathbb{R}^{(T-c)\times(T-c)}$ is the degree matrix, $\mathbf{H}^{(1)} \in \mathbb{R}^{(T-c)\times d_2}$ represents the initial node embeddings, $\mathbf{H}^{(l)} \in \mathbb{R}^{(T-c)\times d_h}$ for $l \geq 1$ represents the node embeddings at layer l, $\mathbf{W}_p^{(1)} \in \mathbb{R}^{d_2 \times d_h}$, $\mathbf{W}_p^{(l)} \in \mathbb{R}^{d_h \times d_h}$ for $l \geq 1$ are learnable weight matrices.

After two graph convolutional layers, the resulting node representations $\mathbf{H}^{(2)} \in \mathbb{R}^{(T-c)\times d_h}$ are then linearly projected using learnable parameters $\mathbf{W}_q \in \mathbb{R}^{d_1 \times d_h}$ and $\mathbf{b}_q \in \mathbb{R}^{d_1}$ to obtain the enhanced topic feature $\mathbf{f}_t^{\mathrm{GCN}} \in \mathbb{R}^{d_1}$ at time step t:

$$\mathbf{f}_t^{\mathrm{GCN}} = \mathbf{W}_q \mathbf{H}^{(2)} + \mathbf{b}_q. \tag{12}$$

3.3 Feature Adaptive Fusion

After obtaining the GRU feature $\mathbf{f}_t^{\mathrm{GRU}} \in \mathbb{R}^{d_1}$ and GCN feature $\mathbf{f}_t^{\mathrm{GCN}} \in \mathbb{R}^{d_1}$, the two features are first merged through a fixed equal-weight linear fusion as the main predictive signal:

$$\mathbf{f}_t^{\mathrm{base}} = 0.5\mathbf{f}_t^{\mathrm{GRU}} + 0.5\mathbf{f}_t^{\mathrm{GCN}}. \tag{13}$$

To account for conflicts between temporal dynamics and structural semantics, an attention-guided multi-expert method is applied to the feature difference:

$$\mathbf{d}_t = \mathbf{f}_t^{\mathrm{GRU}} - \mathbf{f}_t^{\mathrm{GCN}}. \tag{14}$$

Each of the $I \in \mathbb{Z}^+$ experts computes a residual projection of this difference:

$$\mathbf{r}_t^{(i)} = \mathbf{W}_e{}^{(i)}\mathbf{d}_t, \quad i = 1, \ldots, I, \tag{15}$$

where $\mathbf{W}_e{}^{(i)} \in \mathbb{R}^{d_1 \times d_1}$ is the learnable weight matrix for the i-th expert.

Attention weights are then computed over the experts to adaptively obtain a weighted residual:

$$\alpha_t = \mathrm{softmax}\big(\mathbf{W}_b \tanh(\mathbf{W}_a[\mathbf{f}_t^{\mathrm{GRU}}; \mathbf{f}_t^{\mathrm{GCN}}])\big), \tag{16}$$

$$\mathbf{r}_t^{\mathrm{weighted}} = \sum_{i=1}^{I} \alpha_t^{(i)} \mathbf{r}_t^{(i)}, \tag{17}$$

where $\mathbf{W}_a \in \mathbb{R}^{d_1 \times 2d_1}$, $\mathbf{W}_b \in \mathbb{R}^{I \times d_1}$ are learnable parameters and $\alpha_t \in \mathbb{R}^I$ and $\mathbf{r}_t^{\mathrm{weighted}} \in \mathbb{R}^{d_1}$.

Finally, the predicted embedding for the next time step t is obtained by adding the scaled weighted residual to the base fusion:

$$\mathbf{e}_t^{\mathrm{current}} = \mathbf{f}_t^{\mathrm{base}} + \lambda\, \mathbf{r}_t^{\mathrm{weighted}}, \tag{18}$$

where λ is a small scaling factor controlling the influence of the residual and $\mathbf{e}_t^{\mathrm{current}} \in \mathbb{R}^{d_1}$.

3.4 Implementation of LLM-SAN

In the training process, the loss function is defined as the reconstruction error between the predicted post embeddings and the corresponding ground-truth embeddings:

$$loss = \frac{1}{T-c} \sum_{t=c+1}^{T} \|\hat{\mathbf{e}}_t^{\mathrm{temporal}} - \mathbf{e}_t^{\mathrm{temporal}}\|^2, \tag{19}$$

where $\hat{\mathbf{e}}_t^{\mathrm{temporal}}$ is the estimation of $\mathbf{e}_t^{\mathrm{temporal}}$, and $T-c$ denotes the number of future time steps to be predicted. Note that, we here use samples within time steps $t = c+1, \ldots, T$ to predict that within time steps $t = 1, \ldots, c$.

The semantic drift score for a post at time step t is then defined as its prediction error:

$$\epsilon_t = \|\hat{\mathbf{e}}_t^{\mathrm{temporal}} - \mathbf{e}_t^{\mathrm{temporal}}\|^2, \tag{20}$$

where $\epsilon_t \in \mathbb{R}$ quantifies the deviation between the predicted embedding and the actual post embedding.

4 Experiment

In this section, comprehensive experiments are conducted to verify that the proposed model achieves state-of-the-art semantic drift evaluation performance and enables reasonable event stage segmentation.

4.1 Datasets

The experiments are conducted on two benchmark datasets. They are as follows:

DTELS (Zhang et al., 2025) is a Chinese news corpus consisting of temporally ordered articles and reference timelines annotated at multiple granularities, the dataset is available at: `https://github.com/chenlong-clock/DTELS-Bench`. Three levels of temporal granularity are provided: coarse-grained (key milestones), medium-grained (important developments), and fine-grained (detailed evolutionary points). Events containing all three granularities and more than 350 documents are retained. Documents are temporally aligned with reference timelines using TF-IDF similarity and summary coverage, and are labeled according to the earliest matched timeline node (3/2/1/0 from coarse to unmatched). The label magnitude indicates the degree of semantic drift. After preprocessing, 18 event datasets are obtained for semantic drift evaluation.

TwiBot-22 (Feng et al., 2022) is a large-scale Twitter benchmark containing approximately one million users and over 76 million tweets, the dataset is available at: `https://github.com/LuoUndergradXJTU/TwiBot-22`. Event-level datasets are constructed by filtering tweets within predefined time periods using event-specific keywords. Noisy samples are removed, including very short posts, replies, and texts dominated by irrelevant content. Following this procedure, 4 event datasets are extracted for event stage segmentation experiments.

4.2 Baseline Models

To test the effectiveness of LLM-SAN, the following models are considered as baselines:

Rolling LDA (Rieger et al., 2022) detects topic changes by monitoring variations in word distributions across consecutive time windows.

DETM (Dieng et al., 2019) models semantic evolution using trainable topic embeddings and LSTM-based inference to identify topic change points.

ANTM (Rahimi et al., 2024) leverages LLM-generated document embeddings and sliding-window clustering to analyze changes in topic distributions.

All three models are originally designed for coarse-grained, time-slice-level detection. To enable post-level evaluation, they are uniformly adapted: each post is treated as a single time point, text is encoded with a unified LLM, and evaluation metrics are adjusted to measure post-level topic change. Additional modifications ensure compatibility with post-level semantic drift analysis and the temporal dynamics captured by LLM-SAN.

4.3 Parameter Settings and Metrics

The hyperparameter of LLM-SAN are summarized in Table 1. Specifically, the maximum number of stages is set to five.

Moreover, the normalized mean absolute error (MAE) is used to evaluate semantic drift. Given a sequence of predicted semantic drift scores $\hat{y}_i$ and the corresponding multi-granularity timeline labels y_i for a sample, then the normalization is applied by

$$\tilde{y}_i = 3 \times \frac{\hat{y}_i - \min(\hat{y})}{\max(\hat{y}) - \min(\hat{y})}, \tag{21}$$

The predicted value is mapped into the range $[0, 3]$, which is consistent with the granularity labels. The MAE for this sample is then computed as:

$$\mathrm{MAE} = \frac{1}{T-c} \sum_{t=c+1}^{T} |\tilde{y}_i - y_i| \tag{22}$$

where $T - c$ is the number of time steps considered. This provides a unified metric for comparing semantic drift prediction across different models.

Parameter	Value
Number of experts	4
Time window length	12
LLM-based encoder	Qwen3-Embedding-0.6B
LLM-based generator	Qwen2.5-3B-Instruct
Epochs	3000
Hidden_dim	128
Seed	2025
Learning rate	0.001

Table 1: Hyperparameter settings for LLM-SAN.

4.4 Post-level Semantic Drift Evaluation Performance

To evaluate LLM-SAN on post-level semantic drift prediction, experiments are conducted on the DTELS dataset with post-level labels, and compared with baseline models. As shown in Table 2, LLM-SAN achieves the lowest prediction error, reducing the average error by 23.3% over the second-best method. This demonstrates its effectiveness in capturing semantic drift by: (i) Compared with Rolling LDA, LLM-SAN transforms discrete lexical co-occurrences into continuous sentence-level embeddings via the GRU module, capturing accumulated temporal semantic signals across posts and reducing drift errors. (ii) Compared with DETM, LLM-SAN leverages attention-guided multi-expert fusion to adaptively combine the temporal and topic features at multiple granularities, enhancing sparse keywords and post-level semantic signals, improving drift evaluation accuracy. (iii) Compared with ANTM, LLM-SAN explicitly models semantic dependencies between posts via the GCN module and maintains global semantic continuity through LLM-based encoder and generator features, ensuring robust drift detection even in multi-topic or semantically ambiguous stages.

Event ID	Rolling-LDA	DETM	ANTM	LLM-SAN
1	0.7015	1.1403	1.6338	**0.5150**
2	0.6991	1.1012	1.2621	**0.5085**
3	0.5935	1.0975	1.6648	**0.5087**
4	0.4539	1.4679	1.8993	**0.2505**
5	0.6409	1.2552	1.5488	**0.4672**
6	0.8263	1.2469	1.4687	**0.3239**
7	0.5527	1.1801	1.2230	**0.4454**
8	0.5688	1.3101	1.3244	**0.3732**
9	0.6744	1.2501	1.7283	**0.3772**
10	0.7532	1.2115	1.4966	**0.4102**
11	0.5379	1.2801	1.4316	**0.4130**
12	0.8197	1.1848	1.6342	**0.3562**
13	0.6802	0.9499	1.4337	**0.4175**
14	0.5869	1.2271	1.3697	**0.4040**
15	0.6757	1.3419	1.4821	**0.3343**
16	0.6960	0.9867	1.3702	**0.5244**
17	0.5220	1.0791	1.4539	**0.3938**
18	0.7541	1.2466	1.5105	**0.5159**

Table 2: Semantic Drift Evaluation Performence on DTELS.

To further evaluate the methods, the predicted post-level semantic drift values are aggregated into temporal bins of 30 and normalized to $[0, 1)$. Subsequently, a Soft-Peak/Valley Enhancement is applied to this normalized curve. This enhancement operates by accentuating local convexity and concavityconceptually akin to approximating the second derivativewhich elevates peaks and deepens valleys. Fig. 3 shows mean drift values for four selected events, where LLM-SAN predictions align most closely with the drift label, capturing both short-term spikes and long-term trends.

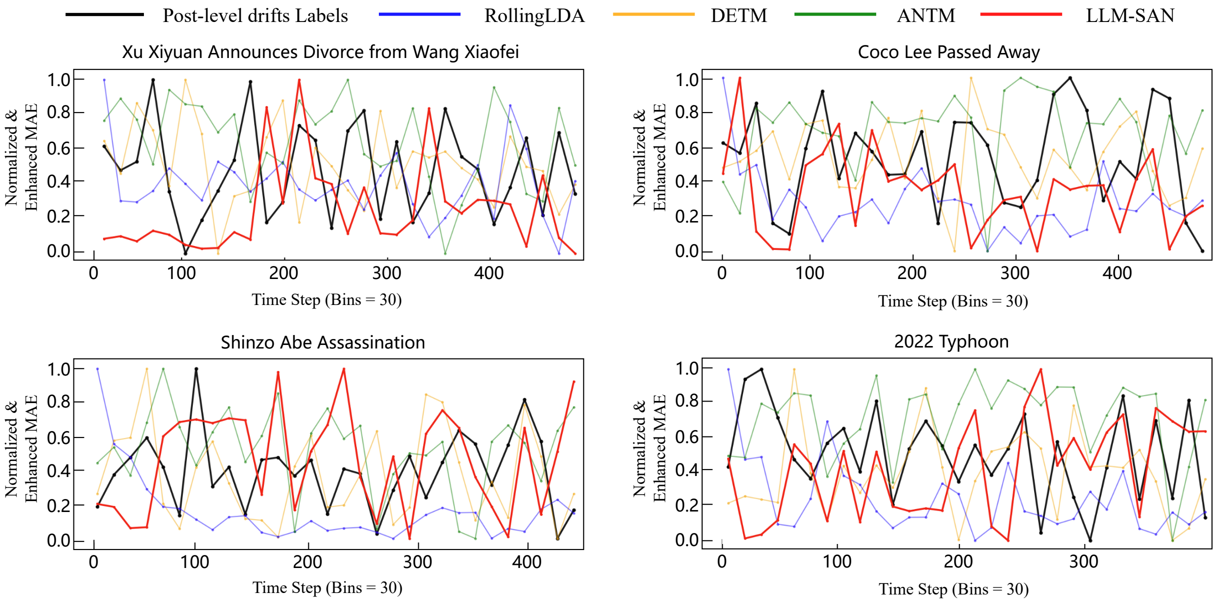

Figure 3: Semantic Drift Trend.

Event ID	Remove GCN	Remove GRU	Remove Expert	LLM-SAN
1	0.5659	0.5839	0.5617	**0.5150**
2	0.5367	0.5017	**0.4694**	0.5085
3	0.5898	0.5310	**0.4918**	0.5087
4	0.3678	0.3276	0.2650	**0.2505**
5	0.5662	0.5328	0.5328	**0.4672**
6	0.3885	0.3424	**0.3141**	0.3239
7	0.4640	0.4398	**0.4080**	0.4454
8	0.5391	0.4769	0.4406	**0.3732**
9	0.4794	0.3937	0.3785	**0.3772**
10	0.5469	0.4805	0.4487	**0.4102**
11	0.4540	0.4687	0.4458	**0.4130**
12	0.5011	0.4401	0.4347	**0.3562**
13	0.4125	0.3947	0.3796	**0.4175**
14	0.5060	0.4198	0.4258	**0.4040**
15	0.4126	0.3665	0.3592	**0.3343**
16	**0.4929**	0.5132	1.3702	0.5244
17	0.4061	0.3971	**0.3823**	0.3938
18	0.6195	0.5080	**0.4714**	0.5159

Table 3: Ablation Experiment Performence on DTELS.

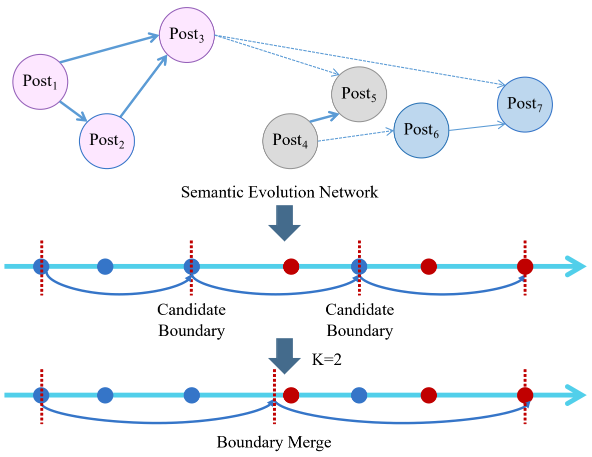

Figure 4: Event Stage Segmentation.

4.5 Ablation Study on Feature Modules and Adaptive Fusion

To evaluate the contribution of LLM-SAN components, ablation experiments are conducted on the DTELS dataset. Variants are created by removing the GRU, GCN, or attention-guided multi-expert fusion module, and compared with the full model, as shown in Table 3. Several observations can be made: (i) The model generally achieves the lowest errors, indicating that single-dimensional features are insufficient to fully capture the post-level semantic evolution. (ii) Incorporating the attention-guided multi-expert fusion allows the model to adaptively reconcile conflicts between temporal (GRU) and topic (GCN) features, substantially reducing errorsfor example, Event 1 and Event 5 drop from 0.5617 and 0.5328 to 0.5150 and 0.4672demonstrating the effectiveness of this adaptive fusion.

4.6 Event Stage Segmentation Experiment

As shown in Fig. 4, event stages are inferred from the post-level semantic drift scores produced by the LLM-SAN model. Following established event lifecycle models (Wang et al., 2020), the segmentation identifies stage boundaries that respect temporal continuity while capturing major fluctuations in posting activity. The procedure selects timesteps with relatively low semantic drift scores and constructs a semantic evolution net-

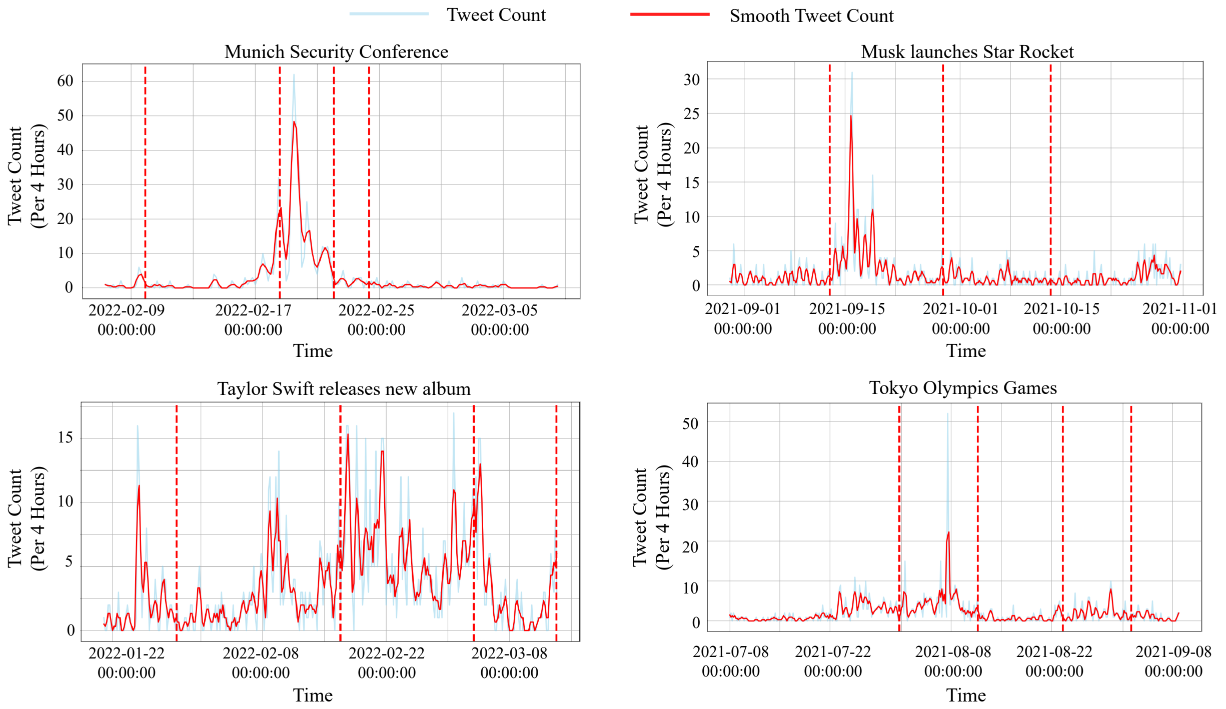

Figure 5: Event Stage Segmentation.

work, where each target post is connected to posts contributing to its prediction. Small connected subgraphs are removed to focus on major propagation patterns. Candidate boundaries are then extracted from the latest posts in retained subgraphs, and an iterative merging process ensures that the total number of boundaries does not exceed a predefined limit $K \in \mathbb{Z}^+$ while maintaining reasonable temporal intervals.

To evaluate the proposed event stage segmentation, experiments are conducted on TwiBot-22 datasets, which offer a large-scale and diverse basis for analyzing event dynamics. In these experiments, the maximum number of stages is set to $K = 5$. Event posts are aggregated in 4-hour windows, smoothed with a three-window moving average, and stage boundaries are visualized as vertical lines on the posting activity curves, as shown in Fig. 5.

Stage boundaries generally occur near small local peaks in low-error regions, marking local maxima or centers of discussion intensity, while adjacent stages are separated by pronounced peaks in post volume, reflecting bursts of activity. These stages do not always correspond to large semantic shifts; instead, they indicate segments where semantic continuity is relatively coherent within each stage but weaker across stages.

5 Conclusion

We propose LLM-SAN for post-level semantic drift evaluation. LLM-SAN integrates a GRU-based module, with inputs from an LLM-based encoder, to capture accumulated temporal semantic signals; a GCN-based module, with inputs from an LLM-based generator, to enhance topic representations and sparse keywords; and an attention-guided multi-expert mechanism to adaptively fuse multi-scale semantic features. Experiments on the DTELS dataset show that LLM-SAN consistently outperforms baseline models in predicting semantic drift, and ablation studies confirm the contribution of each component. Further experiments on the TwiBot-22 dataset demonstrate that LLM-SAN prediction errors can reliably guide event stage segmentation.

Limitations

Despite its effectiveness in post-level semantic drift evaluation and event stage segmentation, LLM-SAN has some limitations: (i) evaluation is limited to Chinese and English, leaving multilingual analysis for future work; (ii) other signals, such as user interactions, are not explicitly modeled and could further enhance performance.

References

Nishant Balepur, Shivam Agarwal, Karthik Venkat Ramanan, Susik Yoon, Diyi Yang, and Jiawei Han. 2023. DynaMiTE: discovering explosive topic evolutions with user guidance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 194–217.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *International Conference on Machine Learning*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Allaa Boutaleb, Jerome Picault, and Guillaume Grosjean. 2024. BERTrend: neural topic modeling for emerging trends detection. In *Proceedings of the Workshop on the Future of Event Detection*, pages 1–17.

Pengfei Cao, Yupu Hao, Yubo Chen, Kang Liu, Jiexin Xu, Huaijun Li, Xiaojian Jiang, and Jun Zhao. 2023. Event ontology completion with hierarchical structure evolution networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 306–320.

Uttam Chauhan and Apurva Shah. 2021. Topic modeling using latent dirichlet allocation: a survey. *ACM Computing Surveys*, 54(7):1–35.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Wentao Fan, Zhiyan Guo, Nizar Bouguila, and Wenjuan Hou. 2021. Clustering-based online news topic detection and tracking through hierarchical bayesian nonparametric models. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2126–2130.

Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, and 1 others. 2022. Twibot-22: towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35:35254–35269.

Rosario Giuseppe Garroppo, Mohamed Ahmed, Saverio Niccolini, and Maurizio Dusi. 2018. A vocabulary for growth: topic modeling of content popularity evolution. *IEEE Transactions on Multimedia*, 20(10):2683–2692.

Wolfgang Gaul and Dominique Vincent. 2017. Evaluation of the evolution of relationships between topics over time. *Advances in Data Analysis and Classification*, 11(1):159–178.

Astha Goyal and Indu Kashyap. 2022. Latent dirichlet allocation-an approach for topic discovery. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing*, volume 1, pages 97–102.

Yanyan He, Mingyu Kang, Duxin Chen, and Wenwu Yu. 2024. Nhce: a neural high-order causal entropy algorithm for disentangling coupling dynamics. *IEEE Transactions on Network Science and Engineering*, 11(6):5930–5942.

Mingyu Kang, Duxin Chen, Ning Meng, Gang Yan, and Wenwu Yu. 2024. Identifying unique spatial-temporal bayesian network without markov equivalence. *IEEE Transactions on Artificial Intelligence*. DOI: 10.1109/TAI.2024.3483188.

Andres Karjus, Richard A Blythe, Simon Kirby, and Kenny Smith. 2020. Quantifying the dynamics of topical fluctuations in language. *Language Dynamics and Change*, 10(1):86–125.

Nozomu Miyamoto, Masaru Isonuma, Sho Takase, Junichiro Mori, and Ichiro Sakata. 2023. Dynamic structured neural topic model with self-attention mechanism. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5916–5930.

Megha Ashok Patil, Sunil Kumar, and Sandeep Kumar. 2025a. Gracod: a disruptive graph-aware drift detection algorithm with a gcn-based time-varying module for concept drift detection in short text streams. *International Journal on Digital Libraries*, 26(4):1–26.

Megha Ashok Patil, Sunil Kumar, and Sandeep Kumar. 2025b. Graphdrift-net: a dynamic graph-based framework for concept drift detection in short unstructured text streams. *The European Physical Journal Plus*, 140(9):884.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282.

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. Antm: aligned neural topic models for exploring evolving topics. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems LVI: Special Issue on Data Management-Principles, Technologies, and Applications*, pages 76–97. Springer.

Jonas Rieger, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch. 2022. Dynamic change detection in topics based on rolling ldas. In *Text2Story@ ECIR*, pages 5–13.

Sayan Unankard and Wanvimol Nadee. 2020. Subevents tracking from social network based on the relationships between topics. In *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on*

Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), pages 1–6.

Kangao Wang, Qizhi Qiu, Mi Wu, and Junan Qiu. 2020. Topic analysis of internet public opinion on natural disasters based on time division. In *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering*, pages 5–10.

Rui Wang, Deyu Zhou, and Yulan He. 2019. Atm: adversarial-neural topic model. *Information Processing & Management*, 56(6):102098.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433.

Yunli Wang and Cyril Goutte. 2018. Real-time change point detection using on-line topic models. In *Proceedings of the 27th international conference on computational linguistics*, pages 2505–2515.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

Chenlong Zhang, Tong Zhou, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Dtels: towards dynamic granularity of timeline summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2682–2703.

Delvin Ce Zhang and Hady Lauw. 2022. Dynamic topic models for temporal document networks. In *International Conference on Machine Learning*, pages 26281–26292.

Sulong Zhou, Pengyu Kan, Qunying Huang, and Janet Silbernagel. 2023. A guided latent dirichlet allocation approach to investigate real-time latent topics of twitter data during hurricane laura. *Journal of Information Science*, 49(2):465–479.