Neural Information Diffusion Prediction with Topic-Aware Attention Network

Hao Wang
Beijing University of Posts and Telecommunications
Beijing, China
wh98@bupt.edu.cn

Cheng Yang*
Beijing University of Posts and Telecommunications
Beijing, China
yangcheng@bupt.edu.cn

Chuan Shi
Beijing University of Posts and Telecommunications
Beijing, China
shichuan@bupt.edu.cn

ABSTRACT

Information diffusion prediction targets on forecasting how information items spread among a set of users. Recently, neural networks have been widely used in modeling information diffusion, owing to the great successes of deep learning. However, in real-world information diffusion scenarios, users are likely to have different behaviors to information items from different topics. Existing neural-based methods failed to model the topic-specific diffusion patterns and dependencies, which have been shown to be useful in conventional non-neural methods. In this paper, we propose Topic-aware Attention Network (TAN) to take advantage of both topic-specific diffusion modeling and deep learning techniques. We jointly model the text content of information items and cascade sequences by incorporating topical context and user/position dependencies into user representations via attention mechanisms. A time-decayed aggregation module is further employed to integrate user representations for cascade representations, which can encode the topic-specific diffusion dependencies independently. Experimental results on diffusion prediction tasks over three realistic cascade datasets show that our model can achieve a relative improvement up to 9% against the best performing baseline in terms of Hits@10.

CCS CONCEPTS

• **Information systems → Data mining**; • **Computing methodologies → Neural networks**.

KEYWORDS

information diffusion, cascade modeling, topic-aware modeling

ACM Reference Format:
Hao Wang, Cheng Yang*, and Chuan Shi. 2021. Neural Information Diffusion Prediction with Topic-Aware Attention Network. In *Proceedings of the 30th ACM Int'l Conf. on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3482374

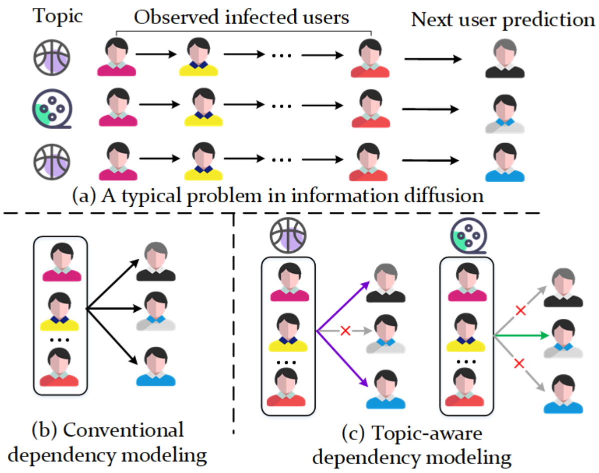

Topic
Observed infected users
Next user prediction
...
(a) A typical problem in information diffusion
(b) Conventional dependency modeling
(c) Topic-aware dependency modeling

Figure 1: An illustrative example of conventional modeling and topic-aware modeling of information diffusion. Conventional modeling captures dependencies entangled from different topics, while topic-aware modeling captures topic-specific dependencies. Here purple arrows indicate the dependencies under the basketball topic, and the green arrow indicates the dependency under the movie topic.

1 INTRODUCTION

Online social platforms such as Twitter and Sina Weibo have attracted millions of users and massive information is diffused among users everyday. The process of information diffusion, also called a cascade, has been analyzed via diffusion pattern and user behavior modeling in many applications, such as popularity prediction [10, 37], epidemiology [26, 36] and personal recommendation [18, 32].

As a popular microscopic cascade prediction task [33], next user prediction [29] has been widely studied in recent years. The problem is formulated as predicting the next infected[1] user of an information item given the time-ordered sequence of previously infected users. With the development of deep learning techniques, some works [6, 16, 34] adopted recurrent neural networks to model cascades by considering information diffusion as sequences of infections and achieved promising performances. Though a cascade is often represented as a sequence of users sorted by infection

*corresponding author.

[1] Conventionally, researchers will use "infect", "activate" or "influence" to characterize the fact that a user interacts with an information item.

CIKM '21, November 1–5, 2021, Virtual Event, Australia.
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8446-9/21/11...$15.00
https://doi.org/10.1145/3459637.3482374

timestamps [17, 35], the real diffusion process [29] is usually non-sequential and depend on unobserved user connection graph. Therefore, other works [29, 31, 33] leveraged attention mechanisms to capture non-sequential long-range diffusion dependencies.

However, existing neural-based methods assumed homogeneous diffusion behaviours and patterns for all information items. This assumption may not hold in the real world. Intuitively, users usually have multiple interests, and their diffusion behaviors could be rather diverse according to the topics of information items. For instance, users are likely to follow and retweet different persons under different topics, and therefore have topic-specific dependencies. Fig. 1 shows an toy example of conventional modeling and topic-aware modeling of information diffusion. Fig. 1(a) presents the typical problem of next infected user prediction in information diffusion analysis. Conventional modeling in Fig. 1(b) usually ignores the text content of diffusing items, and learns mixed dependencies from different topics. In contrast, topic-aware modeling aims to explicit decouple topic-specific diffusion dependencies and thus is able to predict more accurately as shown in Fig. 1(c).

In fact, conventional non-neural methods [1, 12] based on independent cascade (IC) [15] models have demonstrated the advantage of topic-aware modeling, where each information item is recognized as a mixture of multiple topics and the diffusion behaviors under different topics are characterized separately. But these early methods are built on over-strong independent assumptions [33] which limit the generalization performance, and have been shown to be suboptimal by recent deep learning-based methods [27, 33]. To the best of our knowledge, no previous works have suggested a neural-based topic-aware model for capturing different diffusion dependencies from different topics.

In this paper, we propose **T**opic-aware **A**ttention **N**etwork (**TAN**) to benefit from both topic-specific diffusion modeling and deep learning techniques. In specific, we design a novel and effective topic-aware attention mechanism to incorporate the topical context and diffusion history context into user representations for predictions. The topical context enables topic-specific modeling of diffusion patterns, and the diffusion history context can be further decomposed to the user dependency modeling and position dependency modeling. Consequently, we can build multi-topic user representations with context encoded for each user. Then we further integrate user representations for cascade representations by a time-decayed aggregation module. Note that all these modules are motivated by the characteristics of information diffusion. Thus, our proposed TAN can better fit the real-world diffusion data and predict more precisely. Also, the topics are pre-defined in conventional topic-aware models [1, 11] while automatically learned in this work. Experimental results on three public datasets show that our proposed model achieves better performance than state-of-the-art baseline methods on information diffusion prediction. Ablation studies and the analysis of learned topics further demonstrate our effectiveness.

To summarize, the main contributions of this paper are as follows:

• To the best of our knowledge, we are the first work to employ deep learning techniques for topic-aware information diffusion modeling.

• We propose a novel model named TAN to better fit the characteristics of information diffusion. TAN can capture the topic semantic of information text and establish topic-specific diffusion dependencies with attention mechanisms.

• We conduct extensive experiments on three real-world cascade datasets, demonstrating that TAN can significantly improve the prediction performance on next infected user prediction compared with state-of-the-art approaches.

2 RELATED WORK

We group existing microscopic cascade prediction methods into three categories: IC-based methods (popular before 2014), embedding-based (popular during 2014 - 2017) and deep learning based methods (popular after 2017). We summarize related work in Table 1.

Method	Conventional	Topic-Aware
IC model	IC [15], CTIC [22], NetRate [21] NetInf [8],Infopath [9], MMRate [28]	TIC [1], HTM [11] PTC [12]
Embedding	CDK [2], CSDK [2], Embedded IC [3] Inf2vec [7], DNRL [30]	-
Deep Learning	Topo-LSTM [27], CYAN-RNN [29], DeepDiffuse [14], FOREST [34], NDM [33], HiDAN [31], Inf-VAE [24]	**this work**

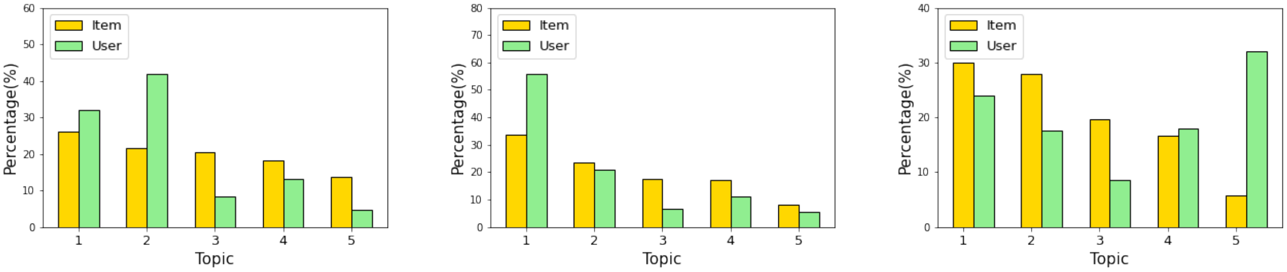
Table 1: Summary of related works.

2.1 IC-based methods

Many cascade diffusion models were based on the assumptions from the fundamental Independent cascade (*i.e.* IC) model [15], which allocates an independent diffusion probability between every user pair. Extensions had been proposed by considering more information, such as continuous timestamps [21, 22] and user profiles [23]. CTIC [22] considered continuous-time information and was able to extracted diffusion paths from sequential observations of infections. NetInf [8] predicted the unobserved diffusion network from temporal correlations and Infopath [9] further inferred the dynamics of the underlying network. MMRate [28] studied multi-aspect diffusion networks since different cascades show various diffusion patterns. A few techniques explored the influence of topic information for cascade modeling. TIC [1] first studied information diffusion from a topic-aware perspective by setting topic-specific probabilities between each user pair. HTM [11] combined hawkes processes with topic modeling to infer diffusion networks. PTC [12] considered users' preferences over topics and designed a preference-enhanced topic-aware cascade model.

2.2 Embedding-based methods

Embedding- based approaches were proposed to take advantage of representation learning techniques. CDK [2] presented an original way to learn diffusion processes by embedding users in a continuous latent space, and CSDK [2] was proposed to additionally take into account diffusion content. Embedded IC [3] computed the diffusion probability of each user pair by a function of their user embeddings instead of directly estimating a real-valued parameter. Inf2vec [7] further combined both the local social influence and global user similarity to learn user embeddings. DNRL [30] learned user representations simultaneously from diffusion sequence and



(a) Weibo

(b) Memetracker

(c) DBLP

Figure 2: Statistical results of the percentages of items and users under each topic on three datasets. The number of LDA topics is set to 5 for all datasets.

social network and was applicable to both diffusion prediction and link prediction tasks. However, neither IC-based methods nor embedding-based methods considered the modeling of sequential information of cascades. Recent works [33] have shown that these models are less effective than deep learning-based models.

2.3 Deep learning-based methods

With the success of deep learning, recurrent neural networks (RNNs) have shown great capability in modeling information diffusion. TopoLSTM [27] extended the standard LSTM model by structuring the hidden states as a directed acyclic graph(DAG) extracted from the social graph. CYAN-RNN [29] and DeepDiffuse [14] employed RNNs and implicitly considered diffusion structures by attention mechanisms. recCTIC [16] proposed a bayesian topological RNN model for capturing tree dependencies. Diffusion-LSTM [13] employed image information to aid the prediction and built a Tree-LSTM model to infer diffusion paths. FOREST [34] extended the GRU model which presented an additional structural context extraction strategy to utilize the underlying social graph information.

Most recently, some attention networks were presented to better capture the diffusion dependencies in cascading sequences. NDM [33] captured dependencies based on multi-head attention mechanism and unified user information with convolution neural networks. HiDAN [31] built hierarchical attention network to jointly capture user dependency and time decay effect. Inf-VAE [24] used a novel expressive co-attentive fusion network to predict the set of all influenced users. HID [38] presented a hierarchical cascade framework by integrating user representation learning and multi-scale modeling. However, to the best of our knowledge, existing neural-based methods rarely took advantage of information text, and did not consider the modeling topic-aware diffusion patterns and user behaviors. Compared with them, TAN also better captures user/position dependencies by sticking to the properties of information diffusion.

3 DATA OBSERVATION

In this section, we will first introduce the datasets studied in this paper. Then we will conduct data observation and investigate the intrinsic relationships between diffusion behaviors and topic semantics. In specific, we will study whether topic preference exists among users and whether users are interested in multiple topics during information diffusion.

Dataset	Weibo	Memetracker	DBLP
# Users	5,000	10,244	4,581
# Links	18,710,854	8,417,276	540,820
# Cascades	27,704	15,521	3,516
Avg. Cascade Length	26.27	15.38	4.37

Table 2: The statistics of datasets.

3.1 Datasets

We collect three real-world cascade datasets containing the text content of diffusing information. Each cascade consists of an ordered-sequence of infections where each infection include both infected user and timestamp. All three datasets are publicly available and have been used in existing work [19, 24, 31] on information diffusion prediction.

Weibo [35] dataset contains the logs of user retweets from Sina Weibo, a Chinese micro-blogging platform. Following the settings of previous works [24], we choose the 5000 most popular users, and the retweeting logs related to these users are selected to construct the dataset. Meanwhile, the text content of each tweet is regarded as its information text.

Memetracker [17] dataset collects a million of Web news article and blog posts from August, 2008 to April, 2009, and track how popular phrases, *i.e.,* memes, spread over the Web. Each meme is considered as an information item and each URL of websites or blogs is considered as a user. We filter out the users who appear less than 60 times in all the reposting logs for data cleaning.

DBLP [19] is a citation network dataset widely used for information diffusion study. The dataset contains paper information of authors, publication time, title and references. For each paper, we extract all the authors who have written or cited it and build a cascade sequence of authors. We consider the title of each paper as its information text.

We randomly sample 80% of cascades for training, 10% for validation and the rest 10% for testing. The statistics of datasets are listed in Table 2. Following [33], two users are assumed to have a link between them if they appear in the same cascade sequence.

3.2 Data Analysis

In this subsection, we will validate our motivation of topic-specific modeling. More specifically, we conduct data observations on the three datasets and further answer the following two questions: *Q1: Does topic preference exist among the users? Q2: Are the users interested in multiple topics?*

To answer the two questions, we first need to figure out the topics among information items. With the help of Latent Dirichlet Allocation (LDA), we can compute topic distribution for each item. In this experiment, we set the number of LDA topics to 5 and categorize each item into the topic with the largest probability based on its topic distribution.

To answer the first question, we present the distributions of items and users under each topic, respectively. For every user, we will count the number of times that she/he interacts with an information item of a specific topic. Then we categorize each user into the topic with the most frequent interactions. Fig. 2 presents the percentages of items and users under each topic on the three datasets. Here the item distribution (golden columns) can be seen as a prior distribution of user distribution (green columns): each user randomly interacts with information items. However, we can see that there are significant differences between item and user distributions on all three datasets. For example, topic 5 on DBLP dataset has the least number of items but the most number of users, which indicates that many users prefers items of topic 5 than those of other topics. Therefore, users do have topic preference on information items during the diffusion.

Dataset	Weibo	Memetracker	DBLP
Top-1	0.439	0.543	0.686
Top-2	0.684	0.805	0.956
Top-3	0.875	0.940	0.990

Table 3: The average ratios that each user's interactions can be covered by the interactions of her/his Top-K (K=1,2,3) favourite topics.

To answer the second question, we conduct a group of statistics to figure out the average number of topics that a user interacts with. Specifically, for every user, we compute the average ratio of (# items she/he interacts with under her/his Top-K favourite topics) to (# all items she/he interacts with). In other words, we are trying to figure out how many percent of each user's interactions can be covered by the interactions of her/his Top-K favourite topics on average. Table 3 presents the statistical results for $K = 1, 2, 3$ on the three datasets. We can see that interactions of a single topic (Top-1) cannot have a good coverage for all three datasets, especially for Weibo where social platform users usually have multiple interests. On the other hand, Top-3 topics on Weibo/Memetracker and Top-2 topics on DBLP can cover around 90% interactions. That is to say, Weibo/Memetracker/DBLP users are mainly interested in 3/3/2 of 5 topics, respectively. From the above analysis, we can conclude that users will be influenced by multiple topics, instead of preferring only one topic.

In summary, these data observations demonstrate the existence of topic preference and multi-topic interest of users, which supports our motivation to model topic-specific diffusion behaviors and dependencies for more accurate predictions.

4 METHOD

In this section, we will start by formalizing the diffusion prediction problem and introducing our embedding strategy to encode user/position/text information into vectors. Then we will propose the topic-aware attention layers, which aim at capturing historical diffusion dependencies and time decay effects in different topics. Finally, our model will predict the next infected users given the multi-topic cascade representations processed by topic-aware attention layers. The overall architecture of our proposed TAN model is shown in Fig. 3.

4.1 Problem Formalization

Given user set U, cascade set C and diffusion message set M, the diffusion sequence of i-th information item in M can be defined as cascade $c_i = \{(u_1^i, t_1^i), (u_2^i, t_2^i) \cdots (u_{|c_i|}^i, t_{|c_i|}^i)\}$, where the tuple (u_j^i, t_j^i) denotes that user u_j^i is infected at time t_j^i and the sequence is ranked by their aa dadinfection timestamps. Following the settings in [29], the diffusion prediction task is to predict the next infected user u_{n+1}^i given diffusion text and previously infected users $\{(u_1^i, t_1^i), \cdots (u_n^i, t_n^i)\}$ in cascade c_i for $n = 1, 2, \cdots, |c_i| - 1$.

4.2 Embedding Layer

User embeddings: To capture users' interests and dependencies in different topics, we employ embedding matrix $M^U \in \mathbb{R}^{|U|\times Kd}$ to encode users, where $|U|$ is the total number of users and K, d are the number of topics and latent dimensions respectively. For each user u_j^i in a cascade sequence $\{(u_1^i, t_1^i), \cdots (u_n^i, t_n^i)\}$, its user embedding is $e_j^i = [e_{j,1}^i, \cdots, e_{j,K}^i] \in \mathbb{R}^{K\times d}$, where $e_{j,k}^i$ is the user embedding in the k-th topic.

Positional embeddings: In order to make use of the sequence order information, we assign a learnable positional embedding $pos_j \in \mathbb{R}^d$ to each position j, where pos_j is shared among all cascades.

Text embeddings: We utilize pretrained language models (*e.g.* BERT [5]) to encode the semantic information of diffusion texts. To measure the topical similarity between user embeddings of a specific topic and text embeddings, we transform the text embeddings encoded by BERT-base [5] $x_i \in \mathbb{R}^{768}$ to $y_i \in \mathbb{R}^d$ via a fully-connected layer:

$$y_i = W_x x_i + b_x, \tag{1}$$

where W_x, b_x are the weight matrix and bias vector.

4.3 Topic-Aware Attention Layer

In this subsection, we will further encode various context information into user representations, and then aggregate them with time-decayed weights to generate cascade representations for each topic.

4.3.1 Enhancing User Representations with Context. Now we will incorporate the topical context and diffusion history context into

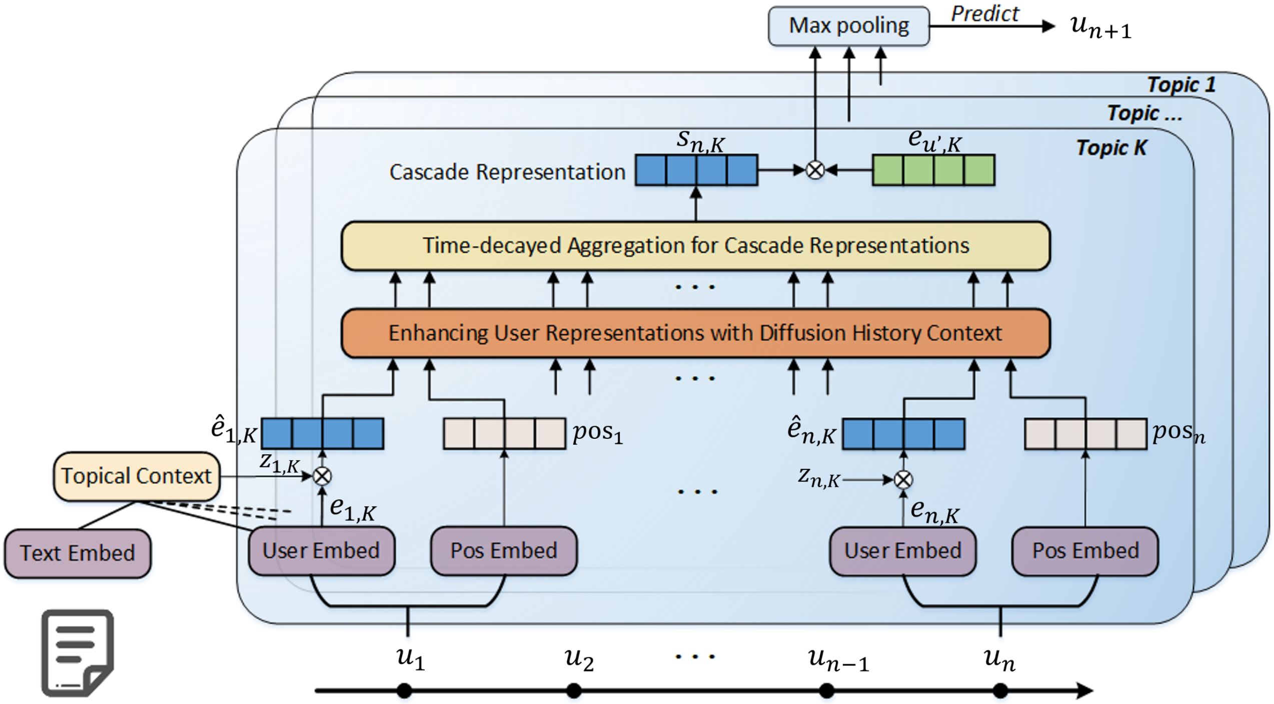


Figure 3: The architecture of our proposed Topic-aware Attention Network (TAN).

multi-topic user representations. The diffusion history context can be further decomposed into user dependency and position dependency. Inspired by multi-head attention [25], we consider a topic as a specific head and perform attention mechanisms in each topic independently to extract user and position dependencies.

1) Topical Context: Given the embedding of diffusion text y_i, we propose to strengthen the user embedding $e^i_{j,k}$ under the k-th topic if there exists a higher similarity between them. In specific, we measure the cosine similarity between $e^i_{j,k}$ and y_i for each topic k, and normalize the similarities by a softmax function:

$$z^i_{j,k} = \frac{exp(\langle y_i, e^i_{j,k} \rangle)}{\sum_{l=0}^{K} exp(\langle y_i, e^i_{j,l} \rangle)}, \tag{2}$$

where $k = 1, 2 \cdots K$ and $z^i_{j,k}$ is the weight for user u^i_j in the k-th topic. Then the user embeddings with topical context is computed as $\hat{e}^i_{j,k} = z^i_{j,k} e^i_{j,k}$. We can see that the k-th embedding $e^i_{j,k}$ having a larger cosine similarity with text embedding will be allocated a larger weight $z^i_{j,k}$ and thus get strengthened.

2) Diffusion History Context: Intuitively, the fact that a user gets infected is usually attributed to the diffusion text and only a few previously infected users in the diffusion sequence. Therefore, the diffusion history context is to extract and characterize the users that are potentially related to the infection of u^i_j. In specific, we propose to employ attention mechanisms for modeling user dependencies, and give more attention weights to such potential users. Formally, the attention score of user dependency between target user u^i_j and its previous user $u^i_m \in \{u^i_1, \ldots, u^i_{j-1}\}$ in topic k is computed as:

$$\alpha^{user}_{jm,k} = (\hat{e}^i_{j,k} W^{tar}_k)(\hat{e}^i_{m,k} W^{pre}_k)^T, \tag{3}$$

where $W^{tar}_k, W^{pre}_k \in \mathbb{R}^{d \times d}$ are topic-specific linear projections for target user and previous user respectively.

Intuitively, we should also pay attention to the source user as well as the newly infected ones. Note that this dependency is independent with specific users, and thus we propose to model position dependency under each topic. Different from previous works which directly sum up predefined positional embeddings and user embeddings, we compute the position dependency scores in a similar way as user dependency modeling. In this way, our method can better capture user-irrelevant position dependencies for better prediction performances.

Then the overall attention score $\alpha^i_{jm,k}$ and weight $w^i_{jm,k}$ between u^i_j and u^i_m can describe the diffusion history context and are computed by

$$\alpha^i_{jm,k} = \frac{1}{\sqrt{2d}} \alpha^{user}_{jm,k} + \frac{1}{\sqrt{2d}} \alpha^{pos}_{jm,k}, \tag{4}$$

$$w^i_{jm,k} = \frac{\exp(\alpha^i_{jm,k})}{\sum_{l=0}^{j-1} \exp(\alpha^i_{jl,k})}, \tag{5}$$

where $\alpha^{pos}_{jm,k}$ is the position dependency score from position m to position j, and $\sqrt{2d}$ is used for re-normalization.

3) Overall multi-topic user representations with context: To take advantage of the topical and diffusion history context, we represent user u^i_j in the k-th topic as the weighted sum of previously infected users:

$$h^i_{j,k} = \hat{e}^i_{j,k} + \sum_{m=0}^{j-1} w^i_{jm,k}\hat{e}^i_{m,k}. \tag{6}$$

Notice that we can also stack multiple layers of the above operations for expressive representations. In this case, the weight $z^i_{j,k}$ for topical context and the position dependency scores $\alpha^{pos}_{jm,k}$ are shared among different layers.

4.3.2 Time-decayed Aggregation for Cascade Representations. After extracting multi-topic representations of users, we need to aggregate them to get cascade representations in multiple topics. We hypothesize that the influence of a user will decay with time, and jointly consider the weights of time decay and diffusion dependencies in Eq. 4.

1) Time-decay effect modeling: Specifically, inspired by [4], we employ non-parametric time decay modeling for each topic. Formally, given the cascade sequence of history infections $\{(u^i_1, t^i_1) \cdots (u^i_n, t^i_n)\}$, we first transform continuous time decay into discrete intervals:

$$f(t^i_n - t^i_j) = l, \ \text{ if } \ \mathbf{t}_{l-1} \le t^i_n - t^i_j < \mathbf{t}_l, \tag{7}$$

where $\mathbf{t}_l$ are defined by splitting the time range $(0, T_{max}]$ into L intervals $\{[0, \mathbf{t}_1) \cdots [\mathbf{t}_{L-1}, T_{max})\}$ and T_{max} is the maximum timestamp in the dataset. Each time interval will have a corresponding learnable weight $\lambda^k_{f(t_n - t_j)}$ for each topic.

2) Computing multi-topic cascade representations: The overall scores for aggregation are calculated by adding an additional term to Eq. (4):

$$\beta^i_{nj,k} = \alpha^i_{nj,k} + \lambda^k_{f(t^i_n - t^i_j)}. \tag{8}$$

Then $\beta^i_{nj,k}$ will be normalized over $j = 1, 2 \cdots n$ via softmax.

Finally, for each topic k, we compute the sum of $h^i_{j,k}$ weighted by $\beta^i_{nj,k}$ over $j = 1, 2 \cdots n$ and employ a point-wise feed-forward network with ReLU activation to endow non-linearity to the model. The output of topic-aware attention layer, *i.e.,* cascade representations, is denoted as $s^i_n = [s^i_{n,1}, s^i_{n,2} \cdots s^i_{n,K}]$.

4.4 Training Objective and Model Details

Given the sequence $\{(u^i_1, t^i_1), \cdots (u^i_n, t^i_n)\}$, the probability of next infected user u^i_{n+1} is parameterized by the similarity between user embedding $e^i_{n+1,k}$ and cascade embedding $s^i_{n,k}$ under the most similar topic. Formally, we compute the likelihood of the cascade interacting with the user u^i_{n+1} as

$$P_\Theta(u^i_{n+1}|s^i_n) = \frac{\max_{k\in\{1,2,\cdots,K\}} \exp(s^i_{n,k} \cdot e^i_{n+1,k})}{\sum_{u'\in U} \max_{k\in\{1,2,\cdots,K\}} \exp(s^i_{n,k} \cdot e^i_{u',k})}, \tag{9}$$

where Θ indicates all the parameters to be learned.

Then our training objective of infected user prediction is to minimize the negative log-likelihood of all users in all the cascades:

$$\mathcal{L}_1 = \sum_{i=1}^{|C|} \sum_{j=1}^{|c_i|-1} -\log P_\Theta(u^i_{j+1}|s^i_j). \tag{10}$$

Besides, we expect that each topic subspace can reflect isolated semantics, and the semantics of different users' embeddings are similar under the same topic. Therefore , we set up K topic prototype embeddings $\{m_k\}_{k=1}^K$ and encourage the user embedding $e^i_{j,k}$ under topic k to be close to the corresponding topic prototype m_k. Formally, we aim to maximize:

$$P_\Theta(k|e^i_{j,k}) = \frac{m_k \cdot e^i_{j,k}}{\sum_{k'=1}^K m_{k'} \cdot e^i_{j,k'}}. \tag{11}$$

Hence, we sum this term over all users as an additional training objective:

$$\mathcal{L}_2 = \sum_{k=1}^{K} \sum_{i=1}^{|C|} \sum_{j=1}^{|c_i|} -\log P_\Theta(k|e^i_{j,k}). \tag{12}$$

The overall training objective function is $\mathcal{L} = \eta\mathcal{L}_1 + (1-\eta)\mathcal{L}_2$, where η is a balance coefficient. We optimize the parameters by gradient descent with Adam optimizer. To avoid unstable training processes, we also apply layer normalization and dropout regularization techniques to user embeddings. Hyperparameter settings will be introduced in next section. Our code is publicly available[2].

5 EXPERIMENTS

In this section, we will conduct experiments on information diffusion prediction task over three public datasets to demonstrate the effectiveness of our proposed model against various baseline methods. We will start by introducing the baseline methods, experimental setting and evaluation metrics. Then we will present experimental results and give further analysis about the evaluation.

5.1 Baselines

We compare the proposed model with a number of state-of-the-art cascade prediction models, which can be roughly classified into four types: IC-based model (i.e., TIC), embedding-based model (i.e., DNRL), RNN-based methods (i.e., CYAN-RNN, DeepDiffuse, FOREST), and attention based methods (i.e., NDM, HIDAN, Inf-VAE).

TIC [1] extends the classic IC model to be topic-aware, and uses an EM approach for estimating the parameters.

CYAN-RNN [29] extends the RNN-based model by a specific attention mechanism for capturing cross-dependence in a cascade.

DeepDiffuse [14] models temporal information and user sequences by temporal point process and LSTM model, and then employs an attention mechanism to obtain cascade representation.

FOREST [34] is also an RNN-based method which builds a structural context extraction strategy to further consider the influence of underlying social graph information.

NDM [33] employs deep learning including convolutional neural network and multi-head attention mechanism to capture user dependencies.

HiDAN [31] is the state-of-the-art attention based sequential model, which adopts a two-level attention mechanism to dynamically capture user dependency and time decay effect.

Inf-VAE [24] presents a variational autoencoder framework to jointly model social homophily and temporal influence, and aims to predict the set of all infected users. We adopt its problem definition to our task by setting the number of infected users to 1.

[2] https://github.com/BUPT-GAMMA/TAN

Model	Weibo				Memetracker				DBLP			
	MRR	*A@10*	*A@50*	*A@100*	*MRR*	*A@10*	*A@50*	*A@100*	*MRR*	*A@10*	*A@50*	*A@100*
TIC	1.04	2.14	7.91	9.87	6.74	12.29	17.08	20.61	9.34	11.56	14.85	17.31
DNRL	2.27	4.37	12.29	19.31	9.14	17.19	30.61	39.24	21.31	24.49	28.96	31.03
CYAN-RNN	1.27	2.31	6.94	9.01	7.53	14.25	25.31	30.51	10.65	15.08	21.88	27.19
DeepDiffuse	1.45	2.97	8.74	14.47	9.07	15.69	31.24	39.37	16.71	20.80	27.64	29.93
FOREST	2.59	4.99	13.95	21.15	12.35	21.87	37.63	46.98	29.68	34.24	38.42	41.30
NDM	1.92	3.62	10.54	16.07	9.73	17.45	31.00	38.24	23.13	26.68	30.79	33.09
HiDAN	2.64	5.24	14.33	21.41	10.73	19.52	34.70	42.78	27.43	32.84	36.54	38.34
Inf-VAE	2.38	4.52	12.82	19.37	9.65	18.22	34.24	43.83	22.43	25.44	29.88	32.15
TAN	**3.07**	**6.13**	**16.33**	**24.17**	**13.28**	**23.92**	**41.06**	**49.84**	**31.31**	**38.92**	**44.91**	**49.43**
Improvement	16.29%	16.98%	13.96%	12.89%	7.53%	9.37%	9.12%	6.09%	5.49%	13.67%	16.89%	19.69%

Table 4: Experimental results on information diffusion prediction. All the metrics are the higher the better.

DNRL [30] explores the correlation between next infected user prediction and link prediction in social network with multi-task predictions.

5.2 Experimental Settings and Metrics

Conventionally, the next infected user prediction task is regarded as an information retrieval problem for evaluation. In other words, information diffusion prediction methods are required to rank all uninfected users by their infection probabilities. Following [31], the prediction performance is evaluated by two widely used ranking metrics: Mean Reciprocal Rank (*MRR*) and Accuracy on top k (*A@k*), where $k = 10, 50, 100$. Larger values of *MRR* and *A@k* indicate better performance.

For hyper-parameter settings, the size of hidden units and user embedding is selected from 64,96,128,160 for all baselines. Other parameters are set according to their original papers. For our model, the size of entire user embedding is set to 160, the topic number is set to $K = 5$, the dimension of topic-specific embedding is $d = \frac{160}{K} = 32$, the number of stacked layers in context encoding is set to 3, and the dropout rate is set to 0.1.

To encode information text, we employ the pretrained language model (*e.g.* BERT [5]) to capture semantic information of text. As demonstrated by [20], the average of context embeddings consistently outperforms the **[CLS]** embedding. Therefore, we use averaging context embeddings as text embedding.

5.3 Main Results

Table 4 presents the overall diffusion prediction performance of all methods on the three datasets. The last row represents the relative improvement of TAN against the best performing baseline method. We have the following observations:

(1) We can find that TAN consistently and significantly outperforms all state-of-the-art baseline methods on all three datasets. As shown in Table 4, the relative improvement over the best performing baseline is at least 5% in terms of *MRR* and *A@k* scores. The improvement on these metrics demonstrates the effectiveness and robustness of our proposed model.

(2) Compared with the traditional topic-aware model TIC, TAN has very significant improvements on both *MRR* and *A@k* metrics: the scores are almost doubled. This indicates the advantage of deep learning techniques for modeling cascade sequences. Compared with the neural models based on RNNs and attention mechanism, the improvements of TAN mostly come from the modeling of topic-specific behaviors, which employs a novel topic-aware attention network to capture the dependencies in each topic. Therefore, the proposed TAN provides a successful modeling to benefit from the advantages of both topic-aware modeling and deep learning techniques.

(3) The improvement is especially impressive on Weibo and DBLP, where the relative improvement over the strongest baselines can go up to 13% in terms of *A@k* scores. The relative improvement gains on Memetracker is around 8%. A possible reason is that information items (memes) spread among websites instead of users in Memetracker dataset, and thus the topic-specific behaviors are less significant. Recall the statistical results in Fig. 2, we can also find that user distributions on Memetracker are also more consistent with item distribution (i.e., priors) than the other two datasets, which indicates weaker topic-specific patterns.

5.4 Analysis of Multi-Interest Users

Since TAN is proposed to model topic-specific diffusion behaviors and dependencies, it should be more suitable for predicting users with multiple interests than baseline methods. In order to prove this ability, we evaluate TAN and all baseline methods only on the appearance of multi-interest users.

First, the topic distribution of information can be obtained by calculating the similarity between learned text embeddings and topic prototypes. Then we compute the topic distribution of a user's interest by averaging the topic distributions of all her/his related information items. Afterward, we select the top 5% users whose topic distributions are the most similar to the uniform distribution (measured by Kullback-Leibler divergence). Hence, the selected users will have multiple interests and it is harder to correctly predict them due to the potential topic-specific dependencies.

As shown in Table 5, we report the results of top 5% multi-interest users on three datasets. For all methods, we can observe a significant decline in all metrics due to the challenge of topic-specific modeling. Compared with baseline methods, our model has a relative improvement of 25% in *MRR* scores and achieves around

Model	Weibo				Memetracker				DBLP			
	MRR	*A@10*	*A@50*	*A@100*	*MRR*	*A@10*	*A@50*	*A@100*	*MRR*	*A@10*	*A@50*	*A@100*
TIC	0.89	1.76	6.41	8.94	4.58	9.41	13.26	15.72	5.27	7.32	10.01	12.36
DNRL	1.64	2.97	8.28	13.41	4.96	10.17	23.64	28.93	12.38	15.25	17.56	19.72
CYAN-RNN	0.74	1.48	4.83	7.62	4.08	8.19	13.62	18.43	5.26	8.74	12.93	15.17
DeepDiffuse	0.95	1.88	6.73	10.54	4.87	11.47	19.08	25.86	9.78	13.22	16.19	18.96
FOREST	2.16	4.25	12.64	18.31	9.24	15.73	26.40	31.42	20.56	25.62	28.91	32.04
NDM	1.31	2.79	8.23	12.36	5.87	10.54	21.27	27.19	14.61	17.26	20.42	24.04
HiDAN	2.04	4.18	12.13	19.27	7.67	14.67	26.08	30.62	17.10	22.74	27.78	29.98
Inf-VAE	1.84	3.67	9.03	15.43	5.04	11.21	25.13	29.84	15.12	18.54	20.17	22.79
TAN	**2.72**	**5.07**	**14.76**	**22.61**	**11.59**	**19.57**	**33.27**	**40.39**	**25.69**	**30.87**	**34.48**	**36.21**
Improvement	25.93%	19.29%	16.77%	17.33%	25.43%	24.41%	26.02%	28.55%	24.95%	20.49%	19.27%	13.01%

Table 5: Experimental results on top 5% multi-interest users. All the metrics are the higher the better.

Model	Weibo				Memetracker				DBLP			
	MRR	*A@10*	*A@50*	*A@100*	*MRR*	*A@10*	*A@50*	*A@100*	*MRR*	*A@10*	*A@50*	*A@100*
RNN+doc	2.04	4.44	12.62	19.01	12.15	21.98	38.46	45.79	27.99	30.81	35.23	37.82
GRU+doc	2.60	5.11	14.20	21.13	12.29	22.00	37.09	45.19	28.73	31.74	35.69	38.06
$\text{TAN}_{\text{K=1}}$	2.70	5.29	14.35	21.47	12.83	22.86	38.79	47.00	28.91	35.63	40.8	42.94
$\text{TAN}_{-\text{pos}}$	2.75	5.49	15.05	22.52	13.54	22.83	40.62	48.97	28.60	33.91	40.15	44.58
$\text{TAN}_{-\text{time}}$	2.95	5.91	15.91	23.77	13.12	23.51	40.84	49.56	29.49	33.66	40.72	45.24
$\text{TAN}_{-\text{text}}$	2.89	5.54	15.21	22.62	12.91	21.87	39.24	47.69	29.45	36.74	40.23	45.07
TAN	**3.07**	**6.13**	**16.33**	**24.17**	**13.28**	**23.92**	**41.06**	**49.84**	**31.31**	**38.92**	**44.91**	**49.43**

Table 6: Experimental results of ablation study.

20% increase in *A*@10. This experiment demonstrates that our proposed model performs well on multi-interest users by considering topic-specific diffusion dependency into cascade modeling.

5.5 Ablation study

In this subsection, we compare several variants of TAN by removing some components. The variant $\text{TAN}_{\text{K=1}}$ removes topic-specific modeling by setting K = 1 to evaluate the benefits from topic-aware attention; $\text{TAN}_{-\text{pos}}$ directly adds positional embeddings to user embeddings for validating the modeling of position dependency; $\text{TAN}_{-\text{time}}$ removes the time-decayed effect to assess its importance; and $\text{TAN}_{-\text{text}}$ ignores the text input and employs uniform topic distributions. In addition, to study the improvement brought by attention mechanisms, we propose two RNN-based methods, *i.e.,* RNN+doc and GRU+doc, which directly concatenate text embeddings encoded by BERT and RNN hidden states, and employ an MLP layer for prediction.

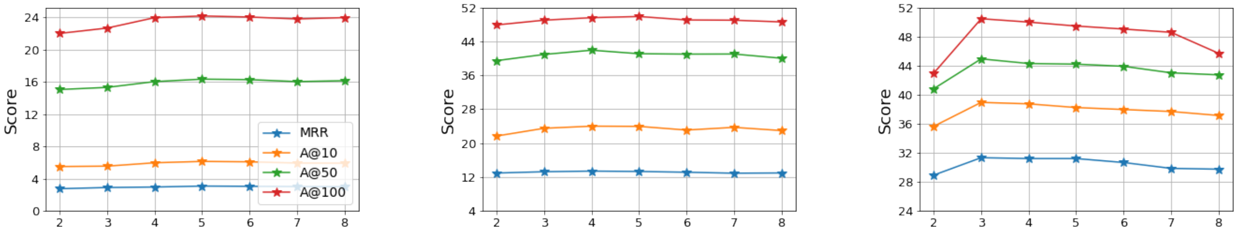
Experimental results of ablation study are shown in Table 6. Comparing to two RNN-based baselines, we can find that our TAN and its variants can achieve better performances, which validates the benefits of the designs in TAN. The performances of TAN against model variants imply the indispensable advantage of these components: topic-specific modeling, position dependency modeling, and time decay aggregation. Among them, topic-specific modeling is the most important module with the largest performance gain. It is worth noting that the model variant $\text{TAN}_{\text{K=1}}$ without topic-specific modeling can still outperforms existing SOTA methods, which indicates that our modeling of context and dependencies is more powerful. These findings demonstrate that the effectiveness of our proposed topic-aware attention mechanism, the primary motivation and contribution of this work.

5.6 Parameter Sensitivity

The proposed TAN model contains several critical hyper-parameters. In this subsection, we will analyse the effect of the following hyper-parameters on prediction performance: 1) the number of topics K, 2) the dimension of topic-specific embedding d and 3) the number of stacked layers in context encoding. We vary each hyper-parameter while keeping others fixed.

The first parameter we evaluated is the number of topics, which we vary from 2 to 8 in this experiment. In Fig. 4, we draw both *MRR* and *A@K* scores to show the effect of the number of topics. For Weibo and Memetracker, the performance grows when the number of topics increases from 2 to 5 and then becomes stable. For DBLP dataset, the model has the best performance with 3 topics, and the scores decrease slightly as the number of topics increases. We hold K=5 for all datasets for convenience.

To evaluate how topic-specific embedding dimension will influence the performance, we fix the number of topics to 5 and try different embedding dimensions. Since the dimension of entire user embedding is computed by $5 \times d$, we only try 16,32,64 to prevent excessive parameters. The comparison results of *MRR* shown in Fig. 5(a) show that the model with $d = 32$ performs better. Moreover, our model does not encounter the overfitting problem when the dimension of embeddings increases from 32 to 64. Thus, 32-dimensional topic-specific embeddings are enough to represent



(a) Weibo

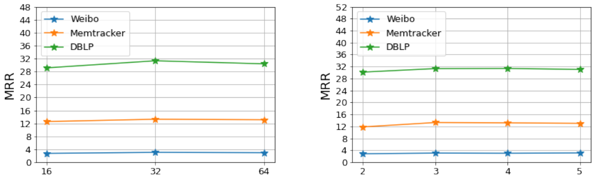
(b) Memetracker

(c) DBLP

Figure 4: Parameter sensitivity analysis of the number of topics.

Topic	Top-frequent words	Avg.depth	Avg.size
1	演唱会 (concert), 音乐 (music), 好听 (pleasant to hear), 作品 (work), 话筒 (microphone), 歌曲 (song)	7.46	24.72
2	法律 (law), 公平 (fairness), 建议 (suggest), 伤害 (hurt), 社会 (society), 保护 (protect)	7.94	28.90
3	好友 (friend), 抽奖 (lottery), 转发 (retweet), 关注 (follow), 分享 (share), 评论 (comment)	7.69	21.84
4	爱心 (love), 期待 (expect), 电影 (movie), 视频 (video), 经典 (classical), 精彩 (wonderful)	8.32	22.32
5	旅行 (travel), 拍摄 (photography), 生活 (life), 好友 (friend), 照片 (photo), 地址 (address)	8.69	32.42

Table 7: The top-frequent words and statistics of diffusion patterns under each learned topic on Weibo dataset.



(a) Embedding dimension

(b) Layer number

Figure 5: Parameter sensitivity analysis of topic-specific embedding dimension and the number of layers.

the semantic space in the dataset and our model is robust to this parameter.

For the number of stacked layers, Fig. 5(b) shows the *MRR* results from 2 to 5 while keeping other hyper-parameters unchanged. A 3-layer architecture yields the best performance, and the results start to decline when we further increase the number of layers because of overfitting.

Based on the experimental results, we can conclude that the proposed model can achieve stable performance, when we tune the hyper-parameters within a reasonable range.

5.7 Analysis of Learned Topics

To further valid the effectiveness of TAN on capturing topic information, we will analyze the semantics of learned topics on Weibo dataset. First, we compute the topic distribution of cascades, and then categorize each diffusion text into the topic with the largest similarity, and conduct statistics about diffusion patterns under each topic. As shown in Table 7, we present the top-6 frequent words, the average depth of the tree-structured diffusion graphs, and the average size of cascades under each topic.

We can see that the semantic meanings and average sizes of different topics are quite diverse from each other. For example, the largest avg.size is about 1.5 times of the smallest one. The 5-th topic about traveling has the largest diffusion depth and size, and thus is the most influential topic. The reason may be that people are willing to retweet photos of beautiful scenes. Therefore, TAN can automatically learn topic semantics with good interpretability.

6 CONCLUSION

In this paper, we propose TAN for topic-specific diffusion modeling. In specific, we aim to jointly model the text content of diffusion items and historical user sequences, and propose topic-aware attentions to capture historical diffusion dependencies and time decay effects in different topics. Compared with conventional topic-aware models, TAN can learn the topics automatically and benefit from the success of deep learning techniques; Compared with existing neural-based models, TAN not only models topic-specific patterns, but also better captures user/position dependencies by sticking to the properties of information diffusion. Experiments on real diffusion datasets demonstrate the effectiveness of our model.

For future works, we plan to consider more information related to diffusion such as the underlying social graph structure, and study multi-modal data fusion method to better capture users' interests and dependencies.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. U20B2045, 62002029, 61772082, 61702296), the Fundamental Research Funds for the Central Universities 2020RC23, and the National Key Research and Development Program of China (2018YFB1402600).

REFERENCES

[1] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2012. Topic-aware Social Influence Propagation Models. In *ICDM*. 81–90.

[2] Simon Bourigault, Cedric Lagnier, Sylvain Lamprier, Ludovic Denoyer, and Patrick Gallinari. 2014. Learning social network embeddings for predicting information diffusion. In *Proceedings of WSDM*. ACM.

[3] Simon Bourigault, Sylvain Lamprier, and Patrick Gallinari. 2016. Representation learning for information diffusion through social networks: an embedded cascade model. In *WSDM*. 573–582.

[4] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. DeepHawkes: Bridging the gap between prediction and understanding of information cascades. In *CIKM*. 1149–1158.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent Marked Temporal Point Processes:Embedding Event History to Vector. In *Proceedings of SIGKDD*. ACM.

[7] Shanshan Feng, Gao Cong, Arijit Khan, Xiucheng Li, Yong Liu, and Yeow Meng Chee. 2018. Inf2vec: Latent Representation Model for Social Influence Embedding. In *ICDE*. 941–952.

[8] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. 2010. Inferring networks of diffusion and influence. In *Proceedings of SIGKDD*. ACM.

[9] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. 2013. Structure and dynamics of information pathways in online media. In *Proceedings of WSDM*. ACM.

[10] Chengcheng Gou, Huawei Shen, Pan Du, Dayong Wu, Yue Liu, and Xueqi Cheng. 2018. Learning sequential features for cascade outbreak prediction. *Knowledge and Information Systems* 57, 3 (2018), 721–739.

[11] Xinran He, Theodoros Rekatsinas, James Foulds, Lise Getoor, and Yan Liu. 2015. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *ICML*. 871–880.

[12] Qingbo Hu, Sihong Xie, Shuyang Lin, Wei Fan, and Philip S Yu. 2015. Frameworks to encode user preferences for inferring topic-sensitive information networks. In *SDM*. 442–450.

[13] Wenjian Hu, Krishna Kumar Singh, Fanyi Xiao, Jinyoung Han, Chen-Nee Chuah, and Yong Jae Lee. 2018. Who Will Share My Image?: Predicting the Content Diffusion Path in Online Social Networks. In *WSDM*. 252–260.

[14] Mohammad Raihanul Islam, Sathappan Muthiah, Bijaya Adhikari, B Aditya Prakash, and Naren Ramakrishnan. 2018. DeepDiffuse: Predicting the'Who'and'When'in Cascades. In *ICDM*. 1055–1060.

[15] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *KDD*. 137–146.

[16] Sylvain Lamprier. 2019. A Recurrent Neural Cascade-based Model for Continuous-Time Diffusion. In *ICML*. 3632–3641.

[17] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD*. 497–506.

[18] Jure Leskovec, Ajit Singh, and Jon Kleinberg. 2006. Patterns of influence in a recommendation network. In *PAKDD*. 380–389.

[19] Cheng Li, Xiaoxiao Guo, and Qiaozhu Mei. 2018. Joint Modeling of Text and Networks for Cascade Prediction.. In *ICWSM*. 640–643.

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks.. In *Proceedings of EMNLP-IJCNLP*.

[21] Manuel Gomez Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. 2014. Uncovering the structure and temporal dynamics of information propagation. *Network Science* 2, 1 (2014), 26–65.

[22] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda. 2009. Learning continuous-time information diffusion model for social behavioral data analysis. In *ACML*. 322–337.

[23] Kazumi Saito, Kouzou Ohara, Yuki Yamagishi, Masahiro Kimura, and Hiroshi Motoda. 2011. Learning diffusion probability based on node attributes in social networks. In *ISMIS*. 153–162.

[24] Aravind Sankar, Xinyang Zhang, Adit Krishnan, and Jiawei Han. 2020. Inf-VAE: A Variational Autoencoder Framework to Integrate Homophily and Influence in Diffusion Prediction. In *WSDM*. 510–518.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[26] Jacco Wallinga and Peter Teunis. 2004. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology* 160, 6 (2004), 509–516.

[27] Jia Wang, Vincent W Zheng, Zemin Liu, and Kevin Chen-Chuan Chang. 2017. Topological recurrent neural network for diffusion prediction. In *ICDM*. 475–484.

[28] Senzhang Wang, Xia Hu, Philip S Yu, and Zhoujun Li. 2014. MMRate: inferring multi-aspect diffusion networks with multi-pattern cascades. In *Proceedings of SIGKDD*. ACM, 1246–1255.

[29] Yongqing Wang, Huawei Shen, Shenghua Liu, Jinhua Gao, and Xueqi Cheng. 2017. Cascade Dynamics Modeling with Attention-based Recurrent Neural Network. In *IJCAI*. 2985–2991.

[30] Zhitao Wang, Chengyao Chen, and Wenjie Li. 2020. Joint Learning of User Representation with Diffusion Sequence and Network Structure. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[31] Zhitao Wang and Wenjie Li. 2019. Hierarchical Diffusion Attention Network.. In *IJCAI*. 3828–3834.

[32] Qitian Wu, Yirui Gao, Xiaofeng Gao, Paul Weng, and Guihai Chen. 2019. Dual Sequential Prediction Models Linking Sequential Recommendation and Information Dissemination. In *Proceedings of SIGKDD*. ACM.

[33] Cheng Yang, Maosong Sun, Haoran Liu, Shiyi Han, Zhiyuan Liu, and Huanbo Luan. 2019. Neural Diffusion Model for Microscopic Cascade Prediction. *IEEE Transactions onKnowledge and Data Engineering* (2019).

[34] Cheng Yang, Jian Tang, Maosong Sun, Ganqu Cui, and Zhiyuan Liu. 2019. Multi-scale Information Diffusion Prediction with Reinforced Recurrent Networks.. In *IJCAI*. 4033–4039.

[35] Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li. 2013. Social influence locality for modeling retweeting behaviors.. In *IJCAI*. 2761–2767.

[36] Liang Zhao, Jiangzhuo Chen, Feng Chen, Fang Jin, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2020. Online flu epidemiological deep modeling on disease contact network. *GeoInformatica* 24, 2 (2020), 443–475.

[37] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD*. ACM, 1513–1522.

[38] Honglu Zhou, Shuyuan Xu, Zuohui Fu, Gerard de Melo, Yongfeng Zhang, and Mubbasir Kapadia. 2020. HID: Hierarchical Multiscale Representation Learning for Information Diffusion. In *IJCAI*.