太原理工大学
TAIYUAN UNIVERSITY OF TECHNOLOGY

Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media

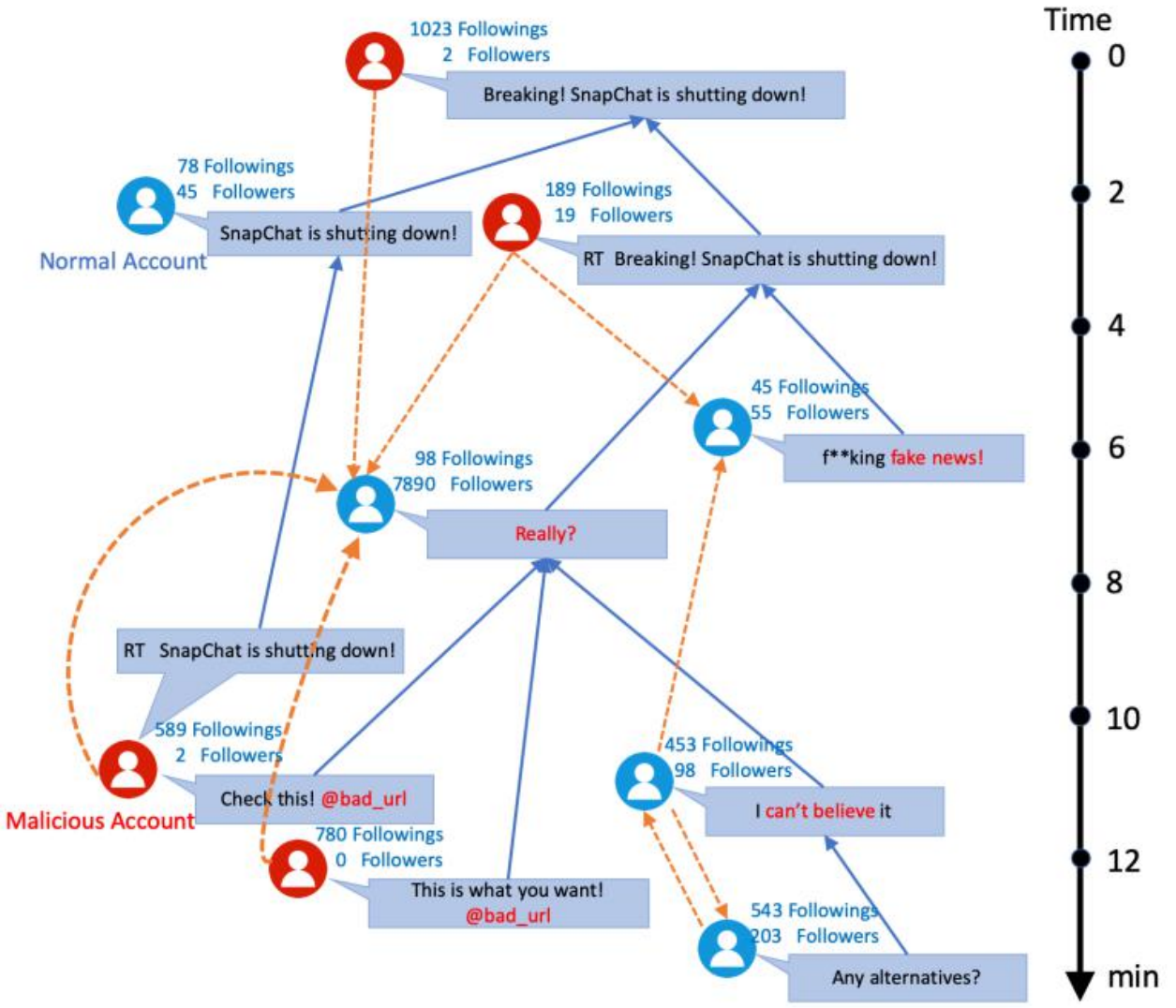
Erxue Min
National Centre for Text Mining,
Department of Computer Science,
The University of Manchester
United Kingdom

Tingyang Xu
Peilin Zhao
Tencent AI Lab
China

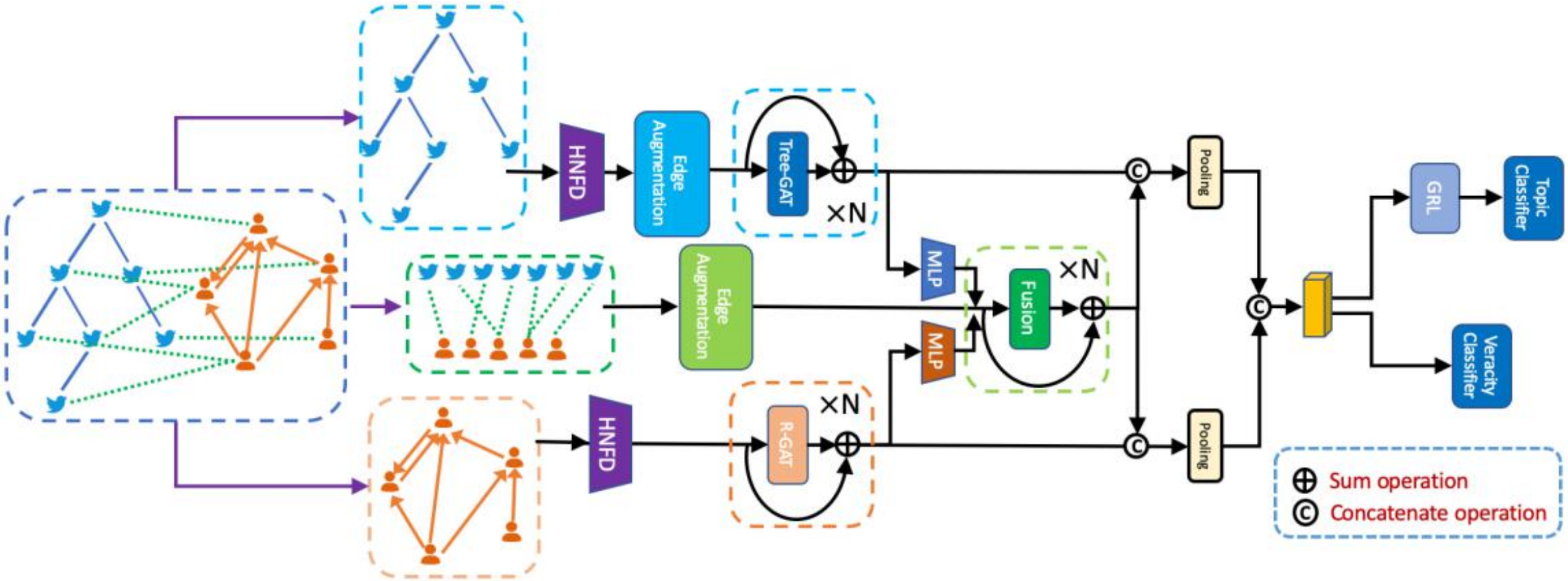
WWW 2022

Zhuomin Chen
2022.09.04

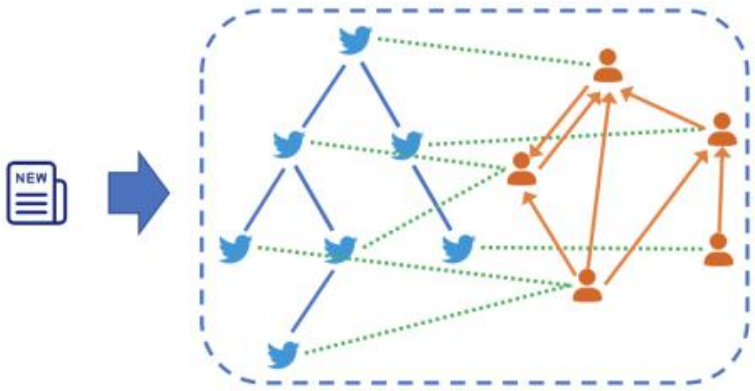
Introduction



Methodology



Problem statement



a fake news dataset with social contexts $\mathbf{D} = \{\mathbf{T}, G^U, G^{UP}\}$

the set of news events $\mathbf{T} = \{T_1, T_2, \ldots T_{|\mathbf{T}|}\}$

the related post set of i-th news event $T_i = \{p_1^i, p_2^i, \ldots p_{M_i}^i, G_i^P\}$

G_i^P is defined as a graph $\{V_i^P, E_i^P\}$ the propagation structure of posts

$V_i^P = \{p_1^i, p_2^i, \ldots p_{M_i}^i\}$ $E_i^P = \{e_{i(st)}^P | s, t = 1., , , .M_i\}$

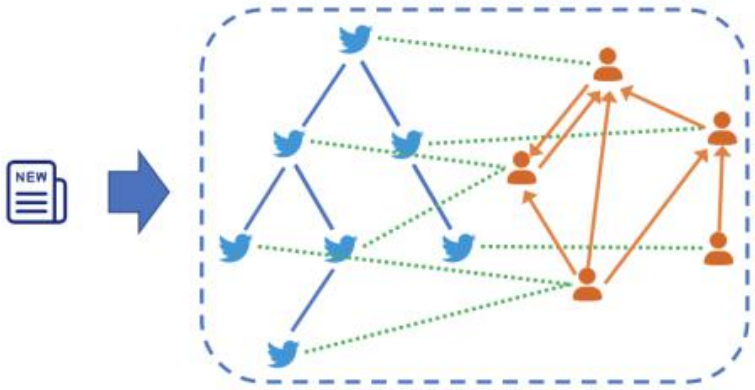
$G^U = \{V^U, E^U\}$ is the user network

$V^U = \{u_1, u_2, \ldots u_N\}$ $E^U = \{e_{st}^U | s, t = 1, 2, , , N\}$

$G^{UP} = \{V^U \cup V^P, E^{UP}\}$ is the bipartite graph between all involved users and all involved posts

$E^{UP} = \{e_{st}^{UP} | s = 1, \ldots N, t = 1, \ldots, M\}$

Problem statement



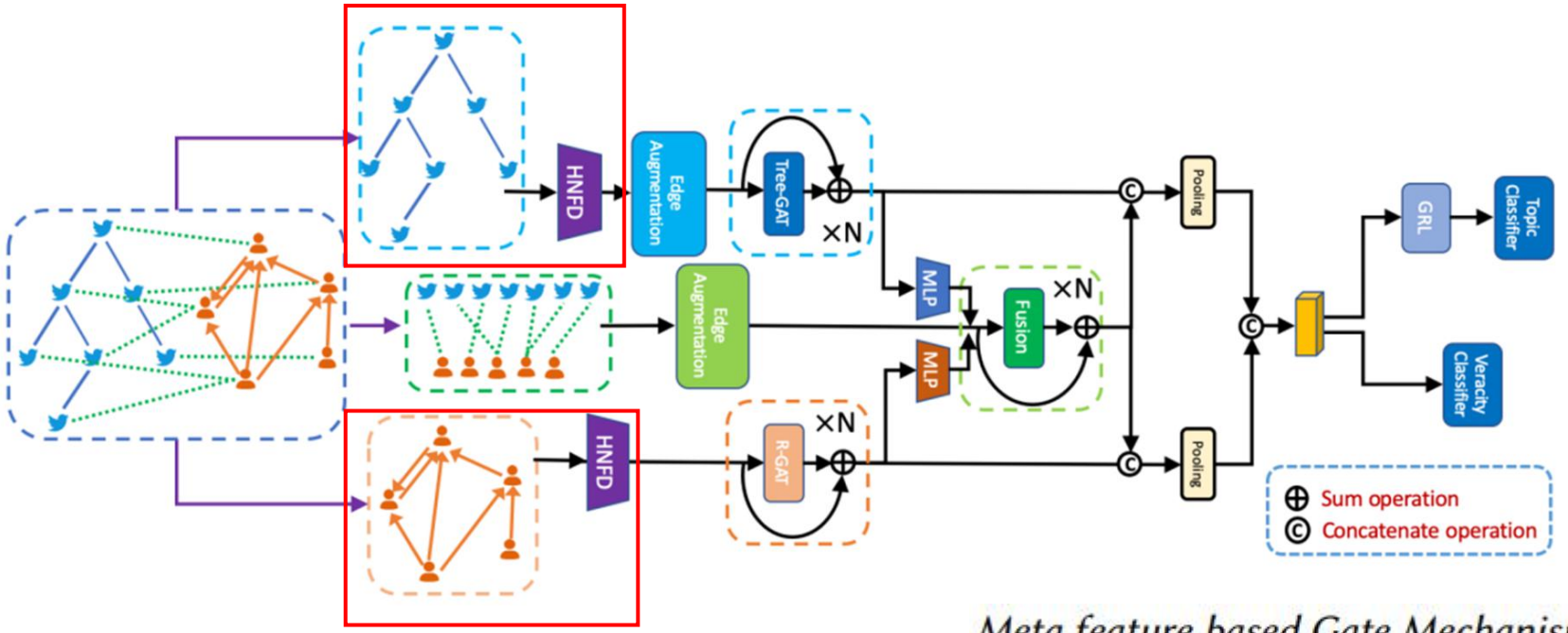
$G^U = \{V^U, E^U\}$ is the user network

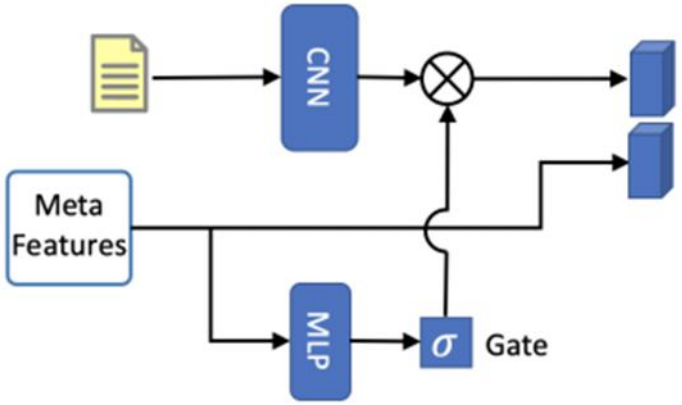
$G^{UP} = \{V^U \cup V^P, E^{UP}\}$ is the bipartite graph between all involved users and all involved posts

G_i^U and G_i^{UP} for each news event T_i

$$T_i = \{p_1^i, p_2^i, \ldots p_{M_i}^i, G_i^P, u_1^i, u_2^i, \ldots u_{N_i}^i, G_i^U, G_i^{UP}\}$$

Hybrid Node Feature Encoder(HNFD)





textual features and meta features,

$$p_j = \{\mathbf{t}_j^p, \mathbf{m}_j^p\}$$

The post meta features $\mathbf{m}^p$ consist of features retweet count, reply count, sentiment score, etc,

$$u_k = \{\mathbf{t}_k^u, \mathbf{m}_k^u\}$$

verified flag, follower count, following count, etc.

Meta feature based Gate Mechanism. $g_j = \sigma(\mathbf{W}^m \mathbf{m}_j + \mathbf{b}^m)$

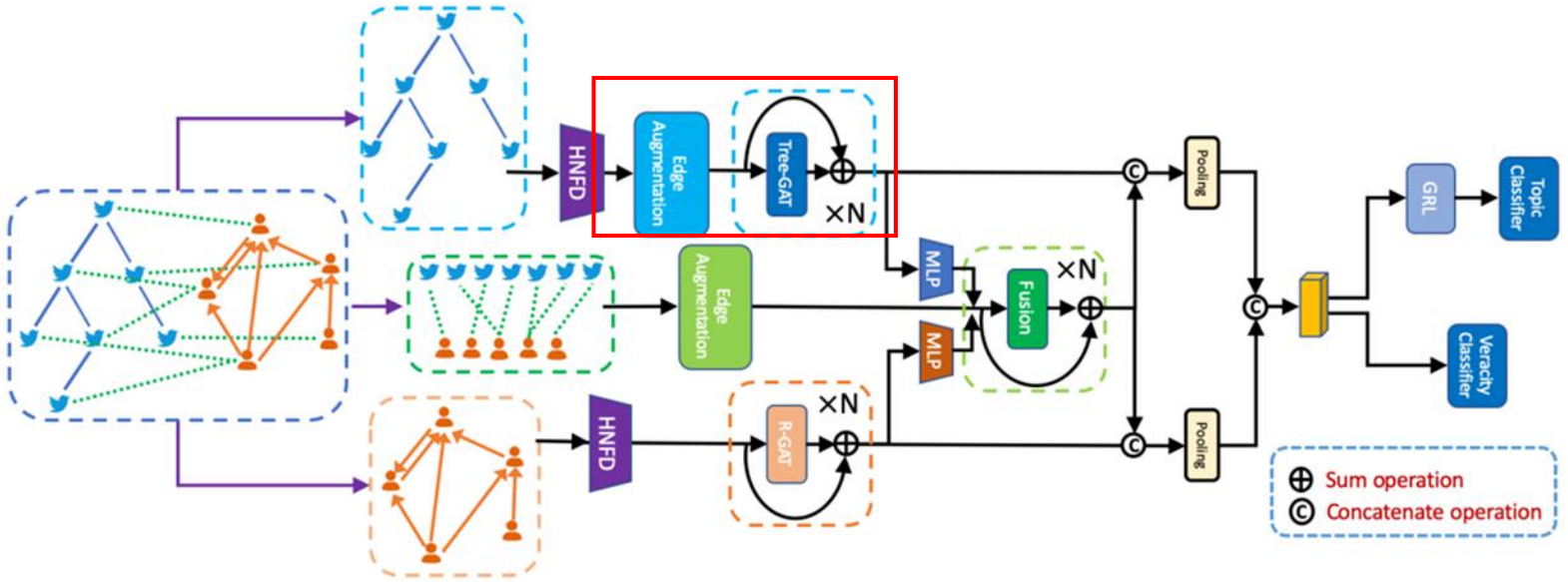
$$\mathbf{n}_j = g_j \mathbf{c}_j \oplus \mathbf{m}_j$$

$\mathbf{c}_j$ be the extracted text embedding for the j-th node.

$$\mathbf{P} = \{\mathbf{h}_1^P, \mathbf{h}_2^P, ..., \mathbf{h}_M^P\}$$

$$\mathbf{U} = \{\mathbf{h}_1^U, \mathbf{h}_2^U, ...\mathbf{h}_N^U\}$$

Post Tree Modeling



$$\alpha_{ij} = \mathrm{Softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})},$$

$$\mathbf{h}_i' = \sigma\Big(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}_d \mathbf{h}_j\Big) + \mathbf{h}_i$$

$$\widehat{\mathbf{P}} = \mathbf{H}^K = \{\widehat{\mathbf{h}}_1^P, \widehat{\mathbf{h}}_2^P, \ldots, \widehat{\mathbf{h}}_M^P\}$$

Edge Augmentation: $A^P_{\mathrm{BU}} = \sum_{d=1}^{d_{\max}} (A^P)^d, A^P_{\mathrm{TD}} = {A^P_{BU}}^{\top},$

$$\widehat{A}^P = A^P_{BU} + A^P_{TD},$$

$d_{\max}$ is the maximum depth of propagation trees

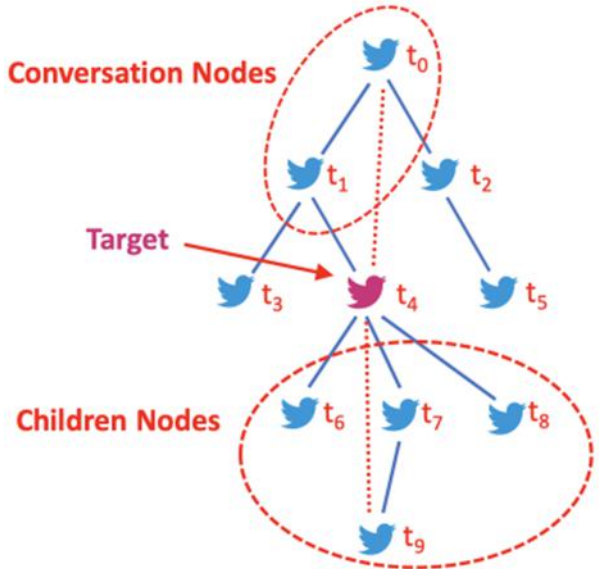
Depth-aware Graph Attention: $e_{ij} = \mathbf{a}^{\top} \mathrm{LeakyReLU}(\mathbf{W} \cdot [\mathbf{h}_i || \mathbf{h}_j] + \mathbf{v}[d(i,j)]),$

(GATv2)

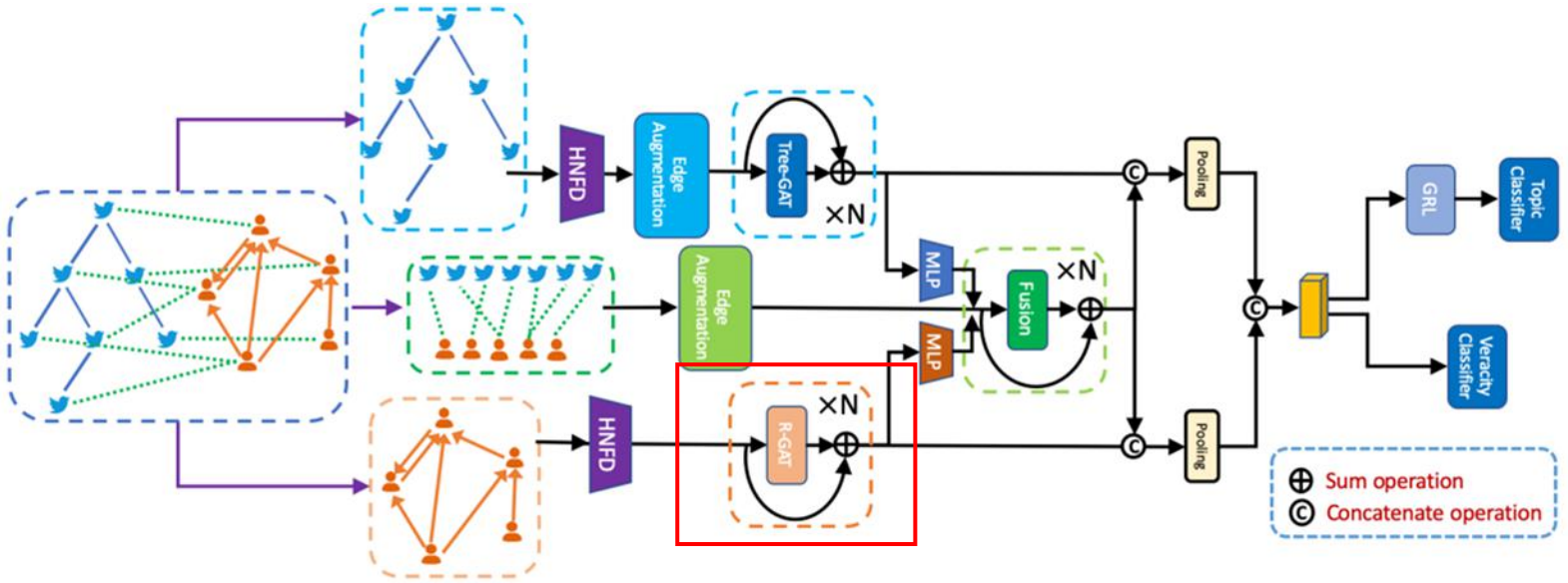
$d(i,j) = d_i - d_j + d_{\max}$ d_i being the depth of i-th node

$d_{\max}$ is the maximum depth of all trees

$\mathbf{v}[d(i,j)] \in \mathbb{R}^d$



User Social Graph Modeling



the neighbours of a user as three groups: only follow relation, only followed, friend (follow and followed)

$$\mathbf{A}^{\text{friend}} = \mathbf{A}^{U} \cdot \mathbf{A}^{U\top}, \mathbf{A}^{\text{follow}} = \mathbf{A}^{U} - \mathbf{A}^{\text{friend}},$$

$$\mathbf{A}^{\text{followed}} = \mathbf{A}^{U\top} - \mathbf{A}^{\text{friend}}.$$

Relational Graph Attention Network (R-GAT): $e_{ij} = \mathbf{a}_{r(i,j)}^{\top}\text{LeakyReLU}(\mathbf{W} \cdot [\mathbf{h}_i || \mathbf{h}_j])$,

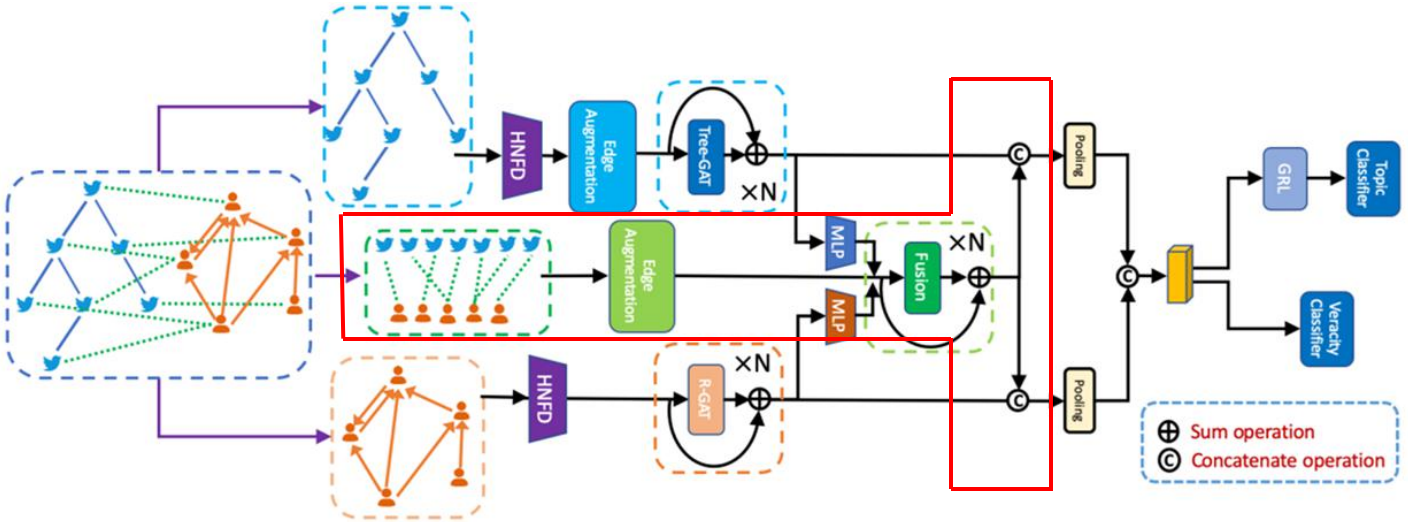
$r(i,j) \in \{0,1,2\}$ denotes the edge type,

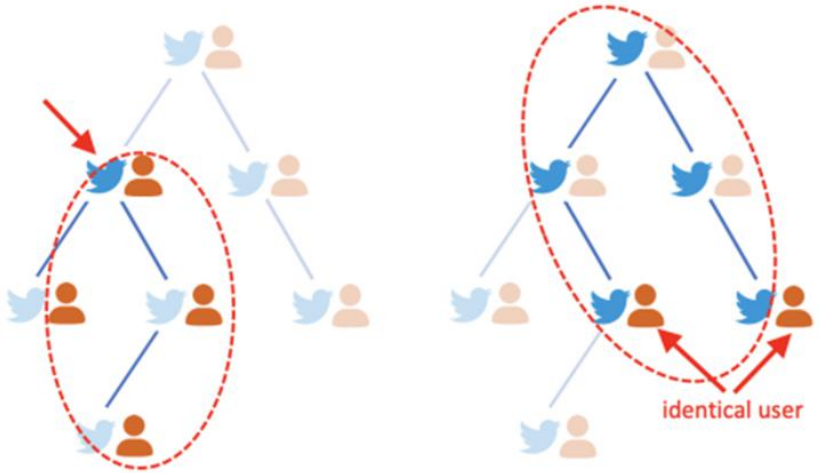
$$\alpha_{ij} = \text{Softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})},$$

$$\mathbf{h}_i' = \sigma\Big(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}_d \mathbf{h}_j\Big) + \mathbf{h}_i$$

$$\widehat{\mathbf{U}} = \{\widehat{\mathbf{h}}_1^U, \widehat{\mathbf{h}}_2^U, \ldots, \widehat{\mathbf{h}}_N^U\}$$

Post-User Interaction





$$\widehat{\mathbf{A}}^{UP} = \mathbf{A}^{UP}\left(\sum_{d=1}^{d_{\max}} (\mathbf{A}^{P})^{d}\right), \quad \widehat{\mathbf{A}}^{UP} \in \mathbb{R}^{N \times M}$$

$$\mathbf{H}^{P} = \mathbf{W}^{P}\widehat{\mathbf{P}}, \mathbf{H}^{U} = \mathbf{W}^{U}\widehat{\mathbf{U}}.$$

$$\mathbf{H} = \text{Concat}(\mathbf{H}^{P}, \mathbf{H}^{U}).$$

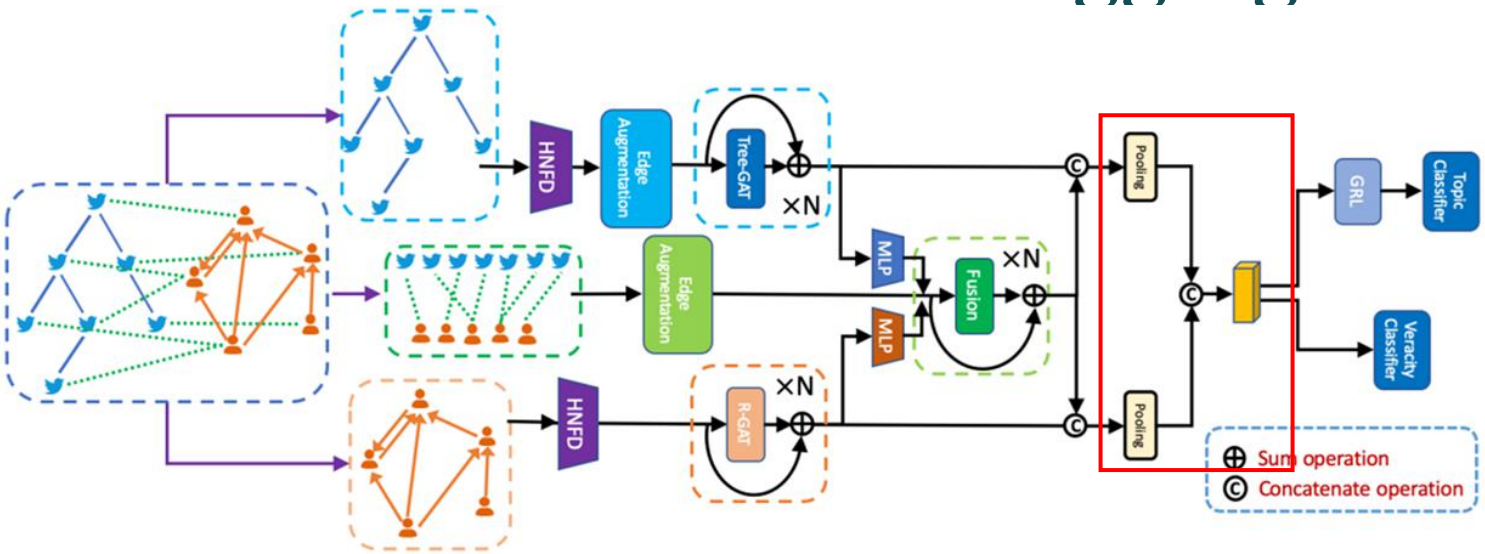
$$\tilde{\mathrm{A}} = \begin{bmatrix} \mathrm{A}^{UP^{T}} & 0 \\ 0 & \mathrm{A}^{UP} \end{bmatrix}$$

$$\mathbf{H}' = \text{GATv2}(\mathbf{H}, \widetilde{\mathbf{A}}) + \mathbf{H}.$$

$$\mathbf{h}_i^{P\prime} = \text{Concat}(\widehat{\mathbf{h}}_i^{P}, \widetilde{\mathbf{h}}_i^{P})$$

$$\mathbf{h}_i^{U\prime} = \text{Concat}(\widehat{\mathbf{h}}_i^{U}, \widetilde{\mathbf{h}}_i^{U})$$

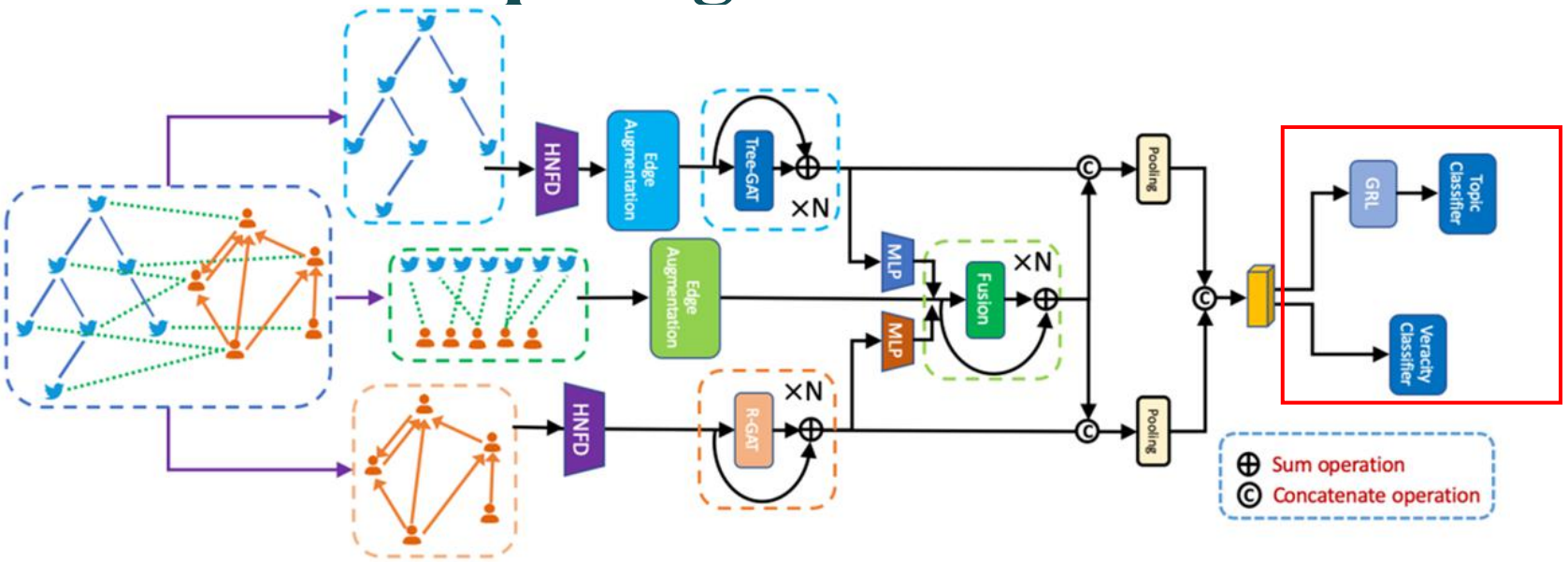
Aggregation



$$\mathbf{r} = \sum_{k=1}^{K} \mathrm{Softmax}(f(\mathbf{h}_k)) \odot \mathbf{h}_k,$$

$$\mathbf{z} = \mathrm{Concat}(\mathbf{p}, \mathbf{u}).$$

Topic-agnostic Fake News Classification



the features from different topics are similar, so that the topic classifier cannot differentiate the topic of the news event.

$$\mathcal{L}(\mathbf{Z}, \mathbf{Y}^V, \mathbf{Y}^C) = \mathcal{L}_V(\mathbf{Z}, \mathbf{Y}^V) + \gamma \mathcal{L}_C(\mathbf{Z}, \mathbf{Y}^C).$$

$$\mathcal{L}_V(\mathbf{Z}, \mathbf{Y}^V) = -\frac{1}{N_t} \sum_{i=1}^{N_t} y_i^V \log(f_V(\mathbf{z}_i))$$ veracity label $y_i^V \in \{F, R\}$ (i.e. Fake, news or Real news)

$$\mathcal{L}_C(\mathbf{Z}, \mathbf{Y}^t) = -\frac{1}{N_t} \sum_{i=1}^{N_t} y_i^C \log(f_C(\mathbf{z}_i))$$ topic label $y_i^C \in$ {Politics, Entertainment, Health, Covid-19, Sryia War}

Gradient Reversal Layer (GRL) $Q_\lambda(x) = x$ with a reversal gradient $\frac{\partial Q_\lambda(x)}{\partial x} = -\lambda I$

Experiments

Table 1: The statistics of the dataset

Topics	Politics		Entertainment		Health		Covid		Syria War	
Labels	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real
News Count	225	1026	2587	8846	590	5120	843	5393	194	2230
Tweet Count (Sum)	51343	140940	128109	504936	75465	695225	33201	285511	9532	227663
Tweet Count (Avg)	228.19	137.37	49.52	57.08	127.91	135.79	39.38	52.94	49.13	102.09
Retweet Count (Sum)	71143	221364	190851	788937	27142	547610	178777	455269	5433	245316
Retweet Count (Avg)	316.19	215.75	73.77	89.19	46	106.96	212.07	84.42	28.01	110.01
Reply Count (Sum)	39342	162108	99362	490452	5682	188730	157835	297559	465	123279
Reply Count (Avg)	174.85	158	38.41	55.44	9.63	36.86	187.23	55.18	2.40	55.28
User Count (Sum)	135338	400815	362195	1504381	91924	1262745	315739	888650	13517	517419
User Count (Avg)	601.50	390.66	140.01	170.06	155.80	246.63	374.54	164.78	69.68	232.03

Experiments

Table 2: Details of the out-of-topic split

ID	Training&Validation set	Testing set
1	Politics, Entertainment, Syria War	Health, Covid-19
2	Health, Covid-19	Politics, Entertainment, Syria War
3	Politics, Entertainment, Health	Covid-19, Sryia War

Table 3: The results of all methods in the in-topic setting.

Methods	Average	
	AUC	F1
SVM	0.7459	0.5210
GRU	0.8539	0.5458
PPC_RNN+CNN	0.8548	0.5419
BiGCN	0.8748	0.5482
PLAN	0.8635	0.5584
FANG	0.8235	0.5084
RGCN	0.8790	0.5930
HGT	0.8856	0.6166
PSIN (-T)	0.9039	0.6213
PSIN	**0.9063**	**0.6267**

Experiments

Table 4: The results of all methods in the out-of-topic setting.

Methods	Average		Split 1		Split 2		Split 3	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
SVM	0.5593	0.1859	0.5737	0.1920	0.5012	0.1273	0.6031	0.2384
GRU	0.6012	0.2118	0.6150	0.2001	0.5298	0.1678	0.6589	0.2675
PPC_R+C	0.6001	0.1984	0.6151	0.1994	0.5344	0.1382	0.6507	0.2576
BiGCN	0.6087	0.2608	0.6201	0.2302	0.5245	0.2086	0.6815	0.3436
PLAN	0.6013	0.1883	0.6133	0.1923	0.5271	0.0283	0.6635	0.3442
FANG	0.6129	0.2371	0.6229	0.2134	0.5381	0.2029	0.6837	0.2949
RGCN	0.6138	0.1949	0.6194	0.2345	0.5400	0.2001	0.6880	0.1501
HGT	0.6147	0.2424	0.6215	0.2357	0.5372	0.2239	0.6913	0.2677
PSIN (-T)	0.6277	0.2693	0.6391	0.2721	0.5469	**0.2459**	0.6971	0.2898
PSIN	**0.6367**	**0.3094**	**0.6571**	**0.2722**	**0.5480**	0.2432	**0.7051**	**0.4120**

Experiments

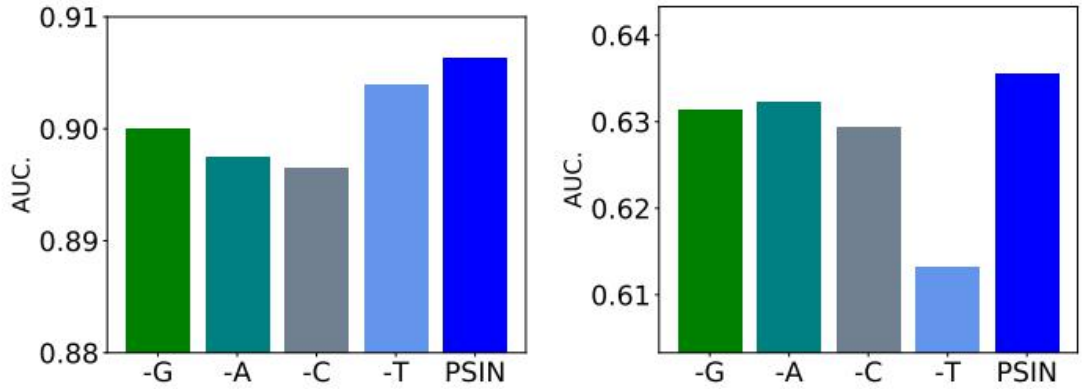
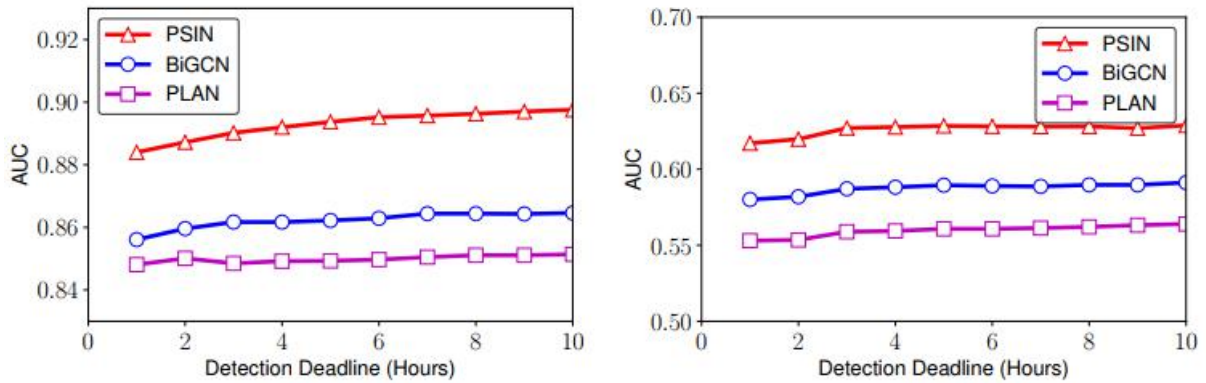


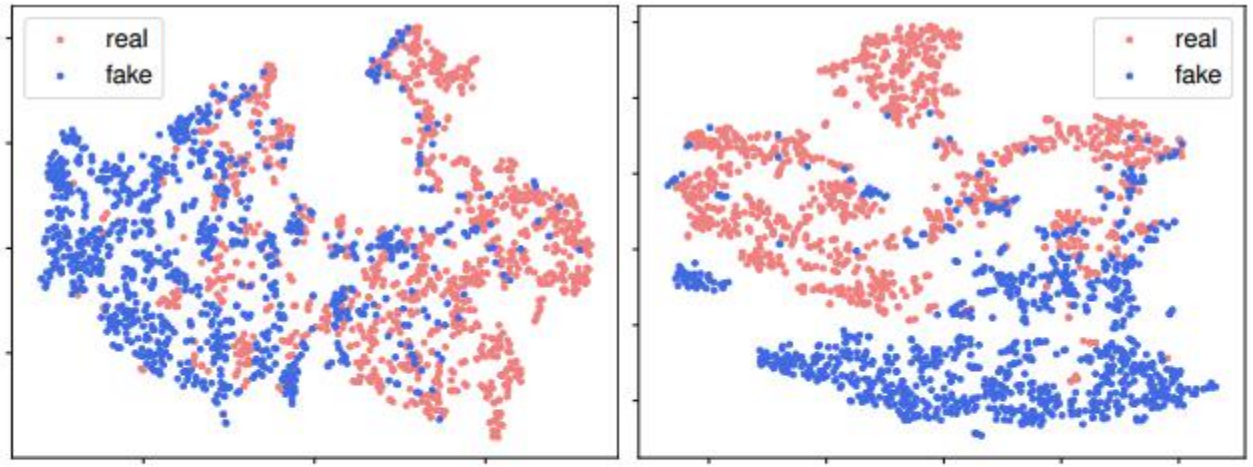
Figure 7: The performance of PSIN and its variants. Left: in-topic performance; Right: out-of-topic performance.



(a) In-topic Split (b) Out-of-topic Split

Figure 8: Results of Early Detection

Experiments



(a) PSIN(-T) (b) PSIN

Figure 9: Visualization of learned feature representations of news events on the testing data.