國際關係學院
University of International Relations

认知博弈背景下社交媒体谣言检测

国际关系学院
李斌阳

认知博弈

- 认知博弈，也称认知对抗、认知域作战，其本质是通过物理域、信息域与认知域的共同行动，夺取人、组织、国家的意志、观念、心理、思维等主导权。

- 大到国家大事层面
  - 俄乌冲突
  - 美国选举

- 小到个体概念层面
  - 东北大学
  - 东大

社交媒体上的认知博弈

- 旨在塑造目标受众的认识、定义、理解事件的宏观框架，并通过宏观框架下的定制信息发送，影响目标受众的行动，继而对现实世界构成重大影响。

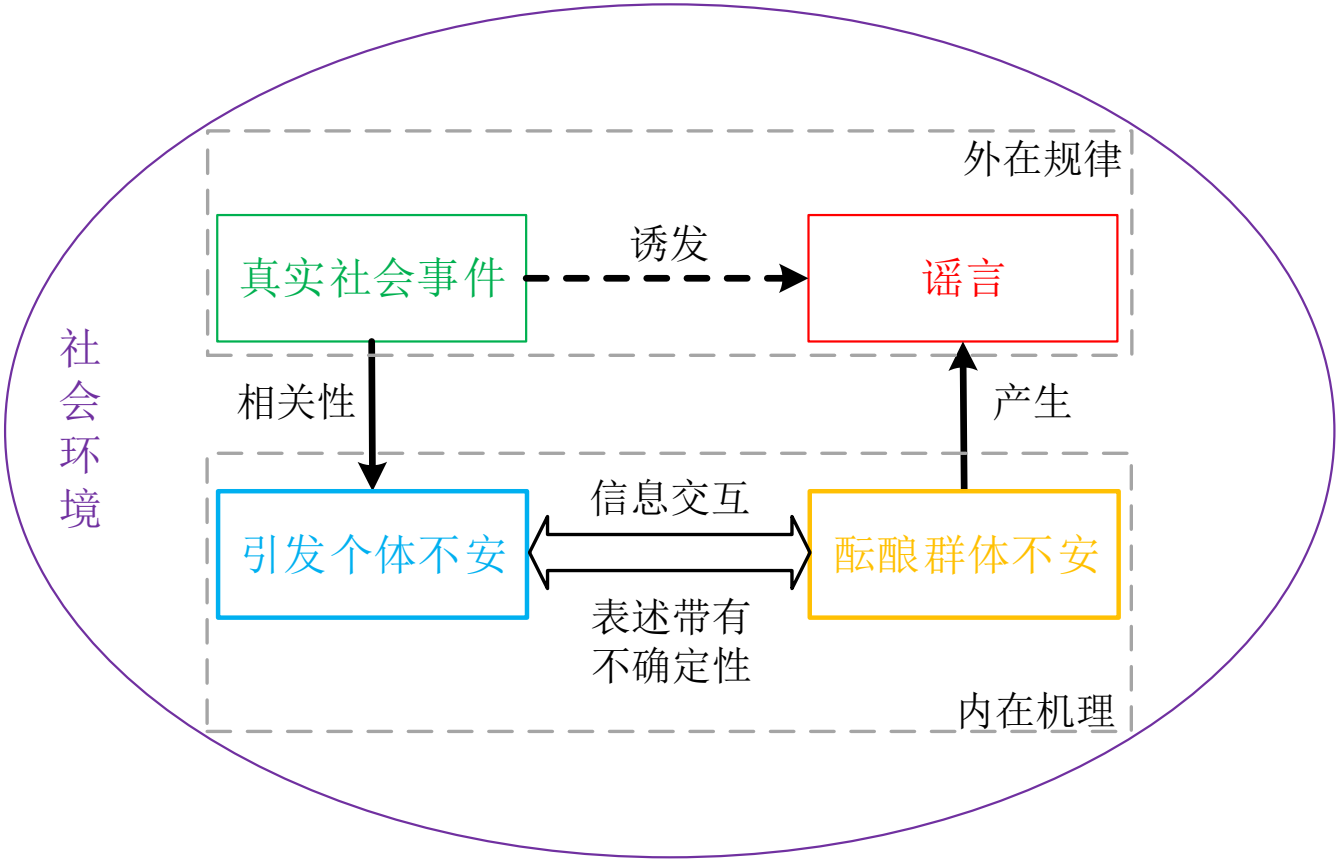
社交媒体谣言检测

- 社交媒体是认知博弈的一个主战场，而在社交媒体上制造谣言则是认知博弈的一种重要手段。

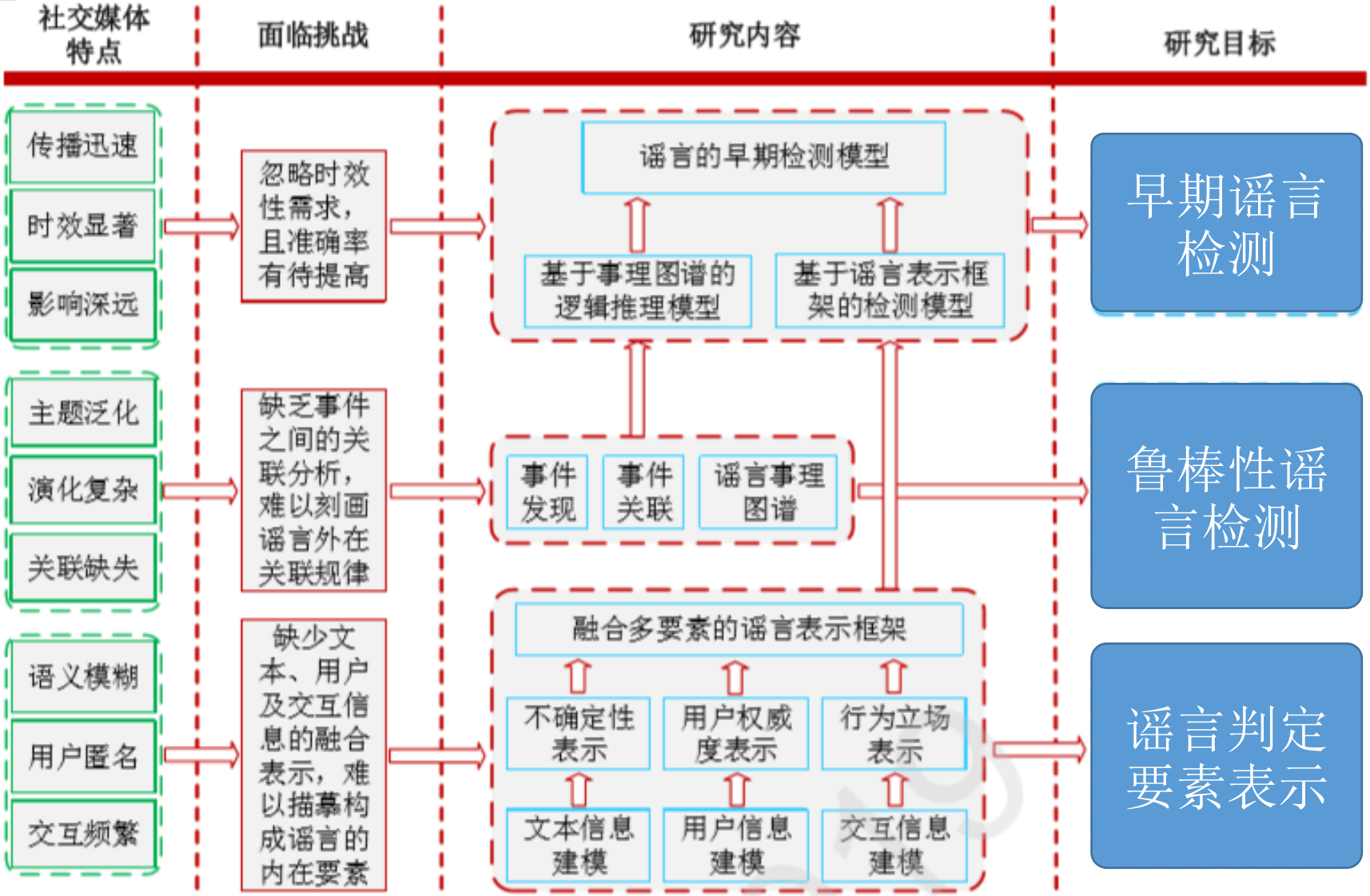
- 谣言的定义
  - 谣言一般是指未经核实的陈述或说明。它往往在一定的社会环境下，伴随某个真实事件，并藉由群体交互中所表达的某种不确定性酝酿发酵而成。

谣言的产生

- 认知科学和社会学的研究表明，谣言的产生是一个伴随某个社会事件的动态过程。



社交媒体的谣言检测


社交媒体特点
面临挑战
研究内容
研究目标
传播迅速
时效显著
影响深远
忽略时效性需求，且准确率有待提高
谣言的早期检测模型
基于事理图谱的逻辑推理模型
基于谣言表示框架的检测模型
早期谣言检测
主题泛化
演化复杂
关联缺失
缺乏事件之间的关联分析，难以刻画谣言外在关联规律
事件发现
事件关联
谣言事理图谱
鲁棒性谣言检测
语义模糊
用户匿名
交互频繁
缺少文本、用户及交互信息的融合表示，难以描摹构成谣言的内在要素
融合多要素的谣言表示框架
不确定性表示
用户权威度表示
行为立场表示
文本信息建模
用户信息建模
交互信息建模
谣言判定要素表示

研究团队近期部分工作

- 谣言检测
  - 谣言早期检测
    - Early Rumor Detection, NAACL, 2019
  - 面向鲁棒性的谣言检测
    - SIFTER: A Framework for Robust Rumor Detection. Trans. Audio Speech Lang. Process, 2021
  - 基于用户行为的谣言检测
    - Social Bot-Aware Graph Neural Network for Early Rumor Detection, COLING, 2022

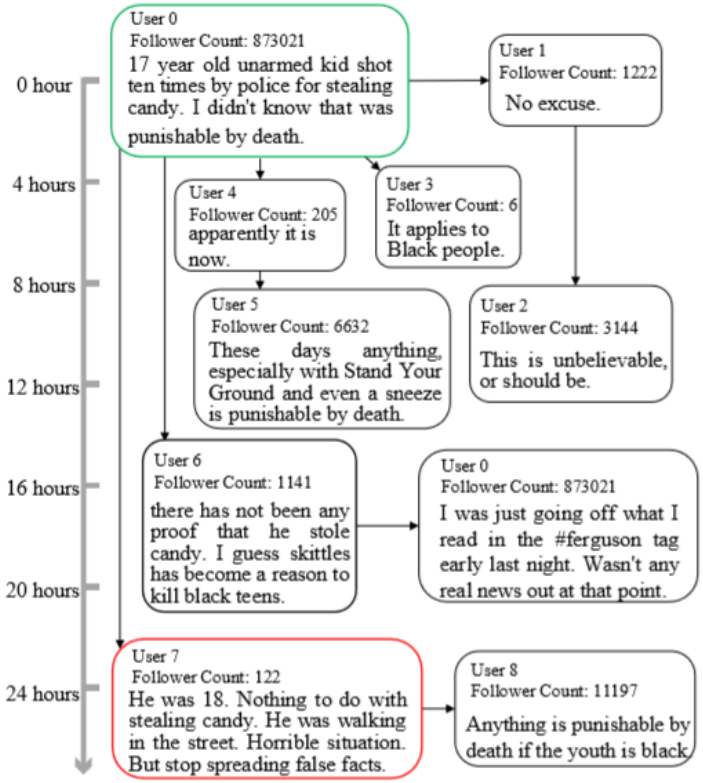
Early Rumor Detection

0 hour

User 0:
17 year old unarmed kid shot ten times by police for stealing candy. I didn't know that was punishable by death

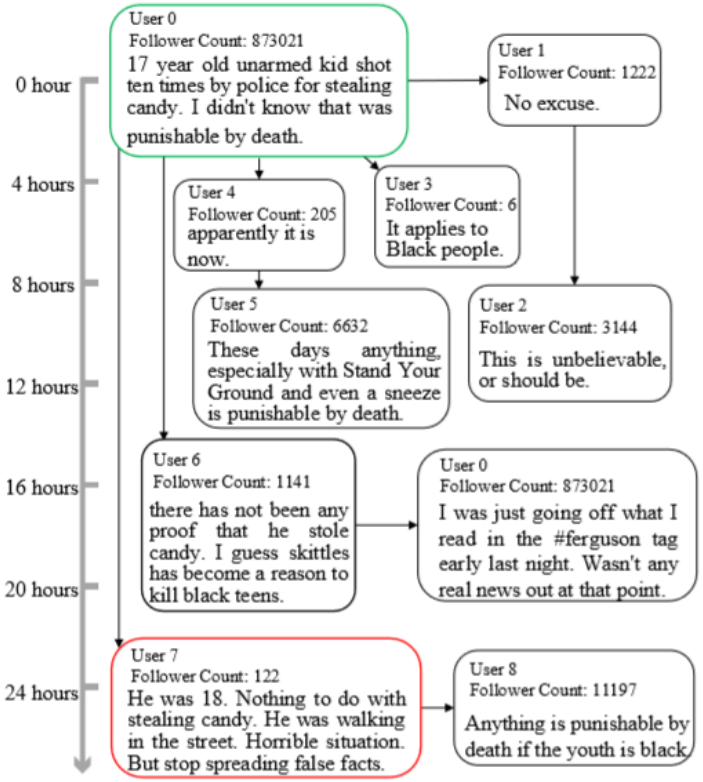
24 hour

User 7:
He was 18. Nothing to do with stealing candy. He was walking in the street. Horrible situation. But stop spreading false facts.



- 早期谣言检测旨在尽早、尽快的检测出网络谣言，从而降低其所引发的危害。

Early Rumor Detection



- 设定某一静态时间点，利用该时间点之前的全部帖子进行判别。
- 例如，设定判别时间为24小时，并将上述帖子看做一篇文档，然后通过提取谣言的相关特征，进行判别。

研究动机和贡献

- 动机：静态时间点难以确定
  - 太早——谣言特征获取不全，判别不准确
  - 太晚——虽然判别准确，但是造成的危害无法挽回

- 贡献
  - 提出一种基于强化学习的早期谣言检测模型。
  - 可以动态确定最优的判别时间点，在保证检测准确率的同时，尽早判别谣言。

任务表述

- 令E表示一个待检测的事件，该事件往往由一组与之相关的帖子构成，其中x_0表示的是源消息，x_1到x_T表示的是针对源消息的跟帖，其中x_T表示的是最后一个回帖。

$$E=\{x_0, x_1, \dots, x_T\}$$

- 早期谣言检测旨在保证准确率的前提下，在更短时间内对该事件进行判别。
  - 换句话说，就是希望利用尽量少的信息实现准确判别。

方法

- 如图所示，early rumor detection 模型包含两个模块，谣言检测模块和判别时间点模块。

Checkpoint Module

Rumor Detection Module

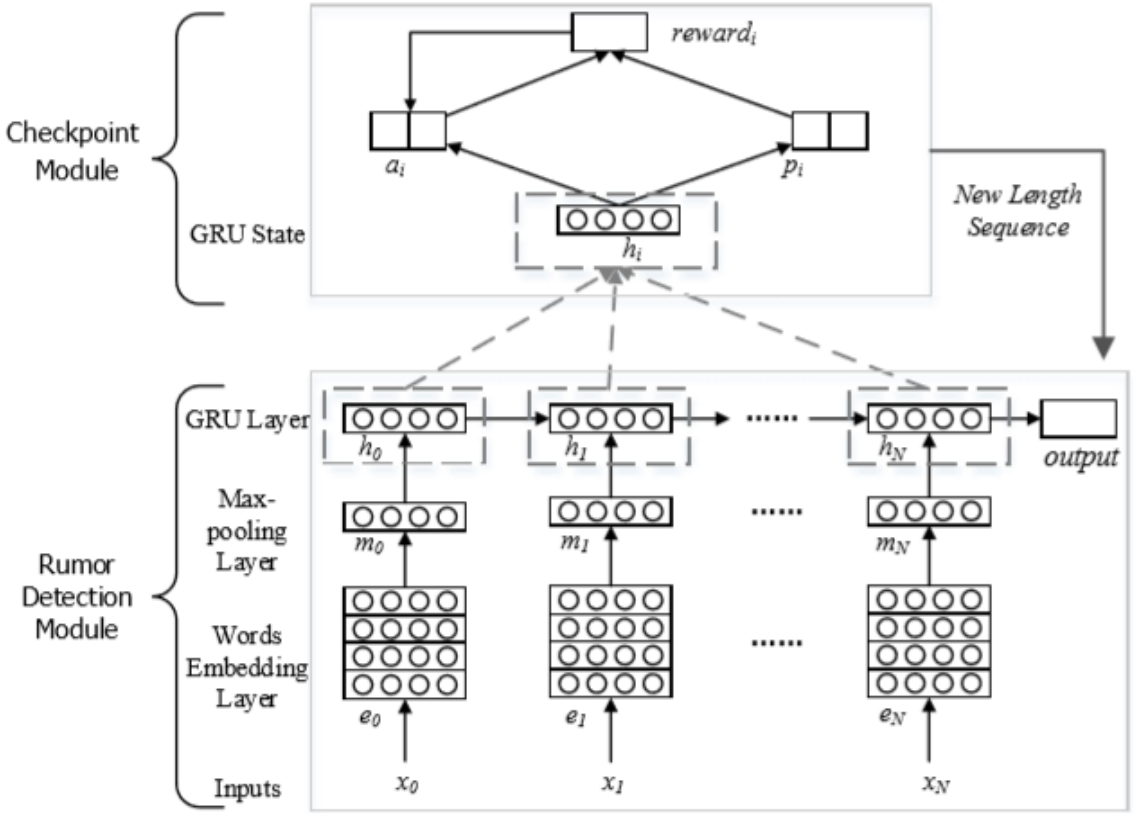


Figure 2: Architecture of ERD.

谣言检测基础模型

- 不同于传统的文本一次性输入方式，ERD根据发布时间对每一个帖子进行排序，并按照顺序进行编码输入检测模型。

$$m_i = \text{maxpool}([\mathbf{W}_m e_i^0; \mathbf{W}_m e_i^1; ...; \mathbf{W}_m e_i^K])$$

$$h_i = \text{GRU}(m_i, h_{i-1})$$

$$p = \text{softmax}(\mathbf{W}_p h_N + b_p)$$

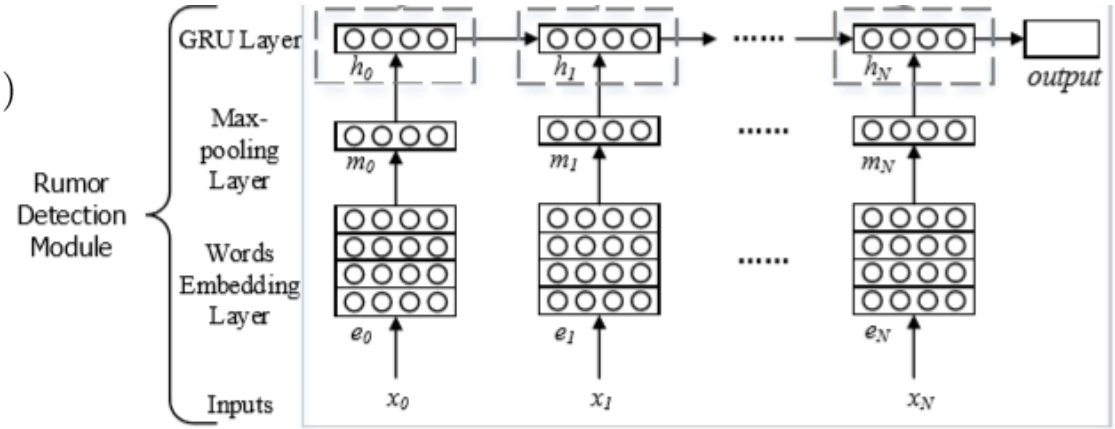


Figure 2: Architecture of ERD.

检测点动态确定模型

$$a_i = \mathbf{W}_a(\mathrm{ReLu}(\mathbf{W}_h h_i + b_h)) + b_a$$

$$a_i \begin{cases} \text{停止读入，进行谣言检测} \\ \text{继续读入新的回贴} \end{cases}$$

$$Q_{i+1}(s,a) = E[r + \gamma \max_{a'} Q_i(s',a')|s,a]$$

$$r_i = \begin{cases} \log M, & \textit{terminate} \text{ with correct prediction} \\ -P, & \textit{terminate} \text{ with incorrect prediction} \\ -\varepsilon, & \textit{continue} \end{cases}$$

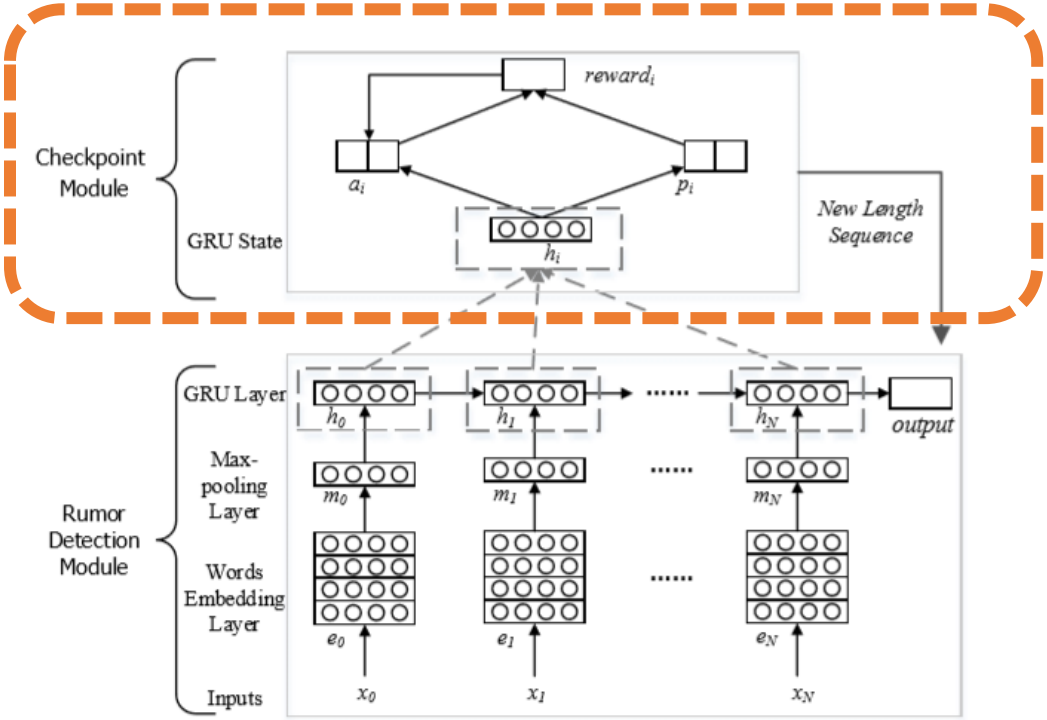


Figure 2: Architecture of ERD.

实验数据

- 在两个谣言数据集上评估模型

Statistics	WEIBO	TWITTER
User#	2,746,818	49,345
Posts#	3,805,656	103,212
Events#	4,664	5,802
Rumors#	2,313	1,972
Non-rumours	2,351	3,830
Avg. hours per event	2,460.7	33.4
Avg. # of posts per event	816	17
Max # of posts per event	59,318	346
Min # of posts per event	10	1

Table 1: Statistics of WEIBO and TWITTER.

实验结果

- 谣言检测准确率结果

Method	Accuracy	Precision	Recall	F1
Baseline	0.724	0.673	0.746	0.707
RNN	0.873	0.816	0.964	0.884
LSTM	0.896	0.846	**0.968**	0.913
GRU-1	0.908	0.871	0.958	0.913
GRU-2	0.910	0.876	0.956	0.914
CSI*	0.953	—	—	0.954
RDM	**0.957**	**0.950**	0.963	**0.957**
ERD	0.933	0.929	0.936	0.932

Table 3: Detection accuracy on WEIBO. '*' denotes values taken from the original publications.

Method	Accuracy	Precision	Recall	F1
Baseline	0.612	0.355	0.465	0.398
RNN	0.785	0.707	0.659	0.682
LSTM	0.796	0.719	0.683	0.701
GRU-1	0.800	0.735	0.685	0.709
GRU-2	0.808	0.741	0.694	0.717
CRF*	—	0.667	0.566	0.607
HMM*	—	—	—	0.524
RDM	**0.873**	0.817	**0.823**	**0.820**
ERD	0.858	**0.843**	0.735	0.785

Table 4: Detection accuracy on TWITTER. '*' denotes values taken from the original publications.

实验结果

- 谣言检测时效性结果

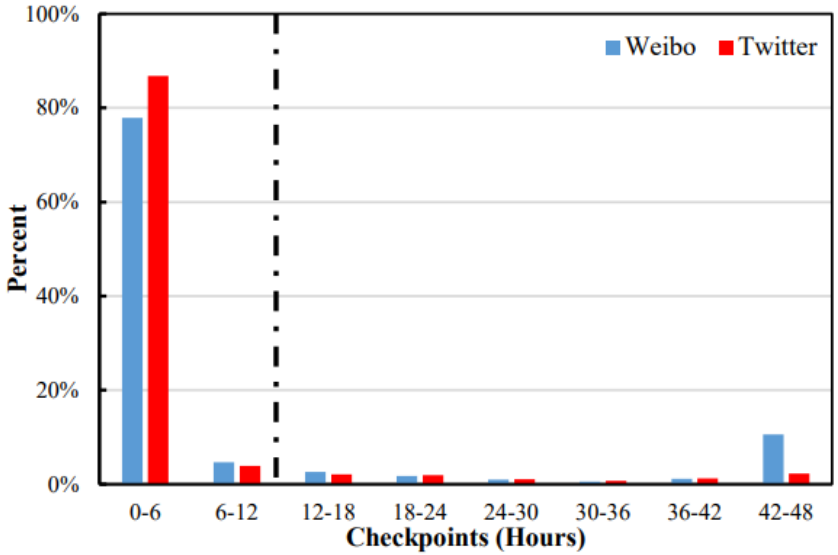


Figure 6: Proportion of events classified by ERD over time. Dashed line indicates the optimal checkpoint (12 hours) for GRU-2.

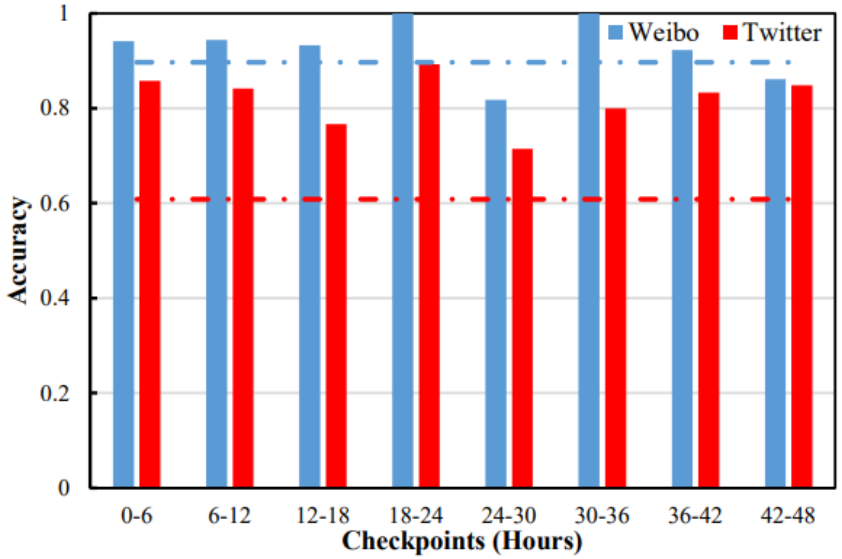


Figure 7: Detection accuracy of ERD over time. Dashed lines indicates GRU-2's accuracies.

案例分析

- 在微博数据集上，在谣言传播1个小时内即可判别真假。

Interval	Salient Words	Translation
18:41 – 18:44	大闸蟹，毒性，激素，有害，吃惊	hairy crabs, toxicity, hormone, harmful, amazed
18:48 – 18:51	大闸蟹，爆出，消息，吃惊，上市	hairy crabs, bursts, message, amazed, on the market
18:51 – 18:59	美食，为何，这样，晕，同城会	delicious food, why, so, dizzy, one city club
18:59 – 19:09	敢吃吗，吃得起，喜欢，惨，偷笑	dare to eat, afford to eat, like, miserable, laughing
19:11 – 19:15	食品安全，真的吗，失望，神马，不能	food safety, really, disappointment, what, cannot
Rumour Detected		
19:34 – 19:49	是不是，大闸蟹，吃不成，疑问，围观	is it, hairy crabs, cannot eat, doubt, look around

Table 5: Case study of a rumour on WEIBO.

SIFTER: A Framework for Robust Rumor Detection

- 在进行谣言检测建模的时候，鲁棒性是一个非常重要的指标，然而现有模型鲁棒性较差，这主要体现在以下两个方面：

预测结果的一致性

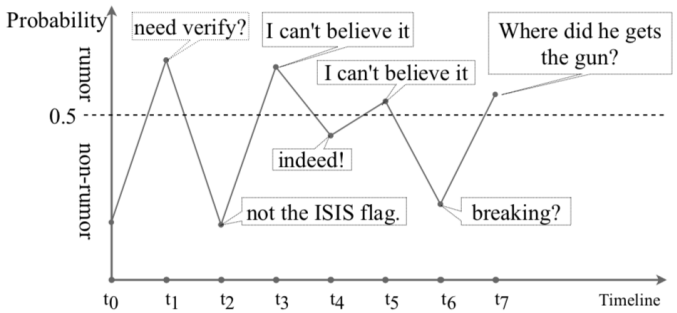


Figure 2: An illustration of prediction inconsistency. At different time points, the model outputs different predictions for an identical event. Sometimes, the prediction of the model even changes at the nearest two timepoints. As a consequence, the reliability of the model is not promised.

跨领域适应性

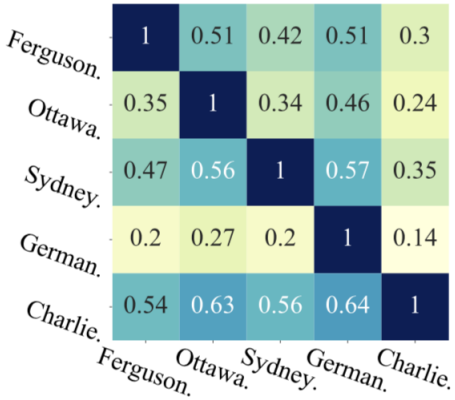
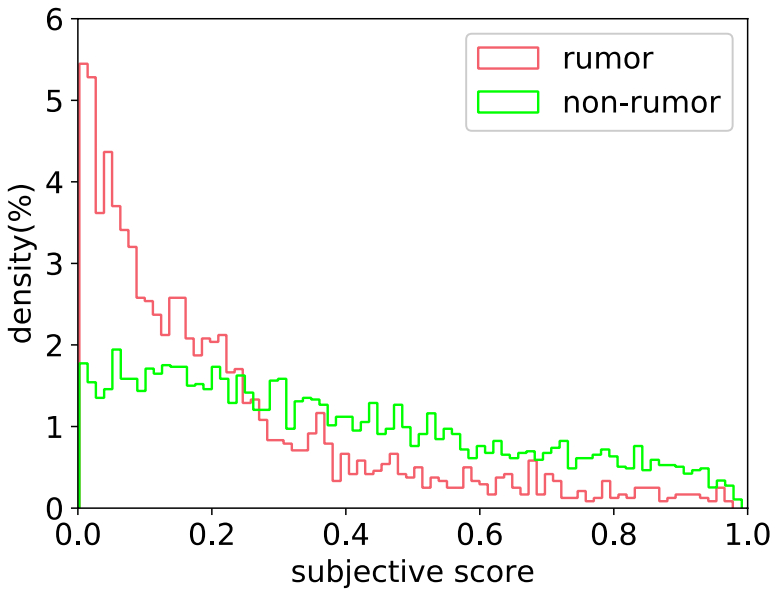


Figure 4: Vacabulary overlap between domains.

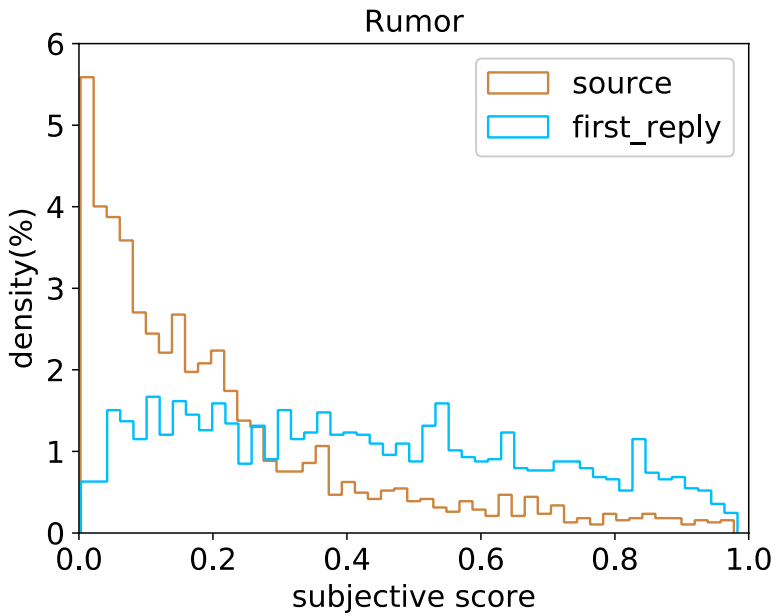
研究动机

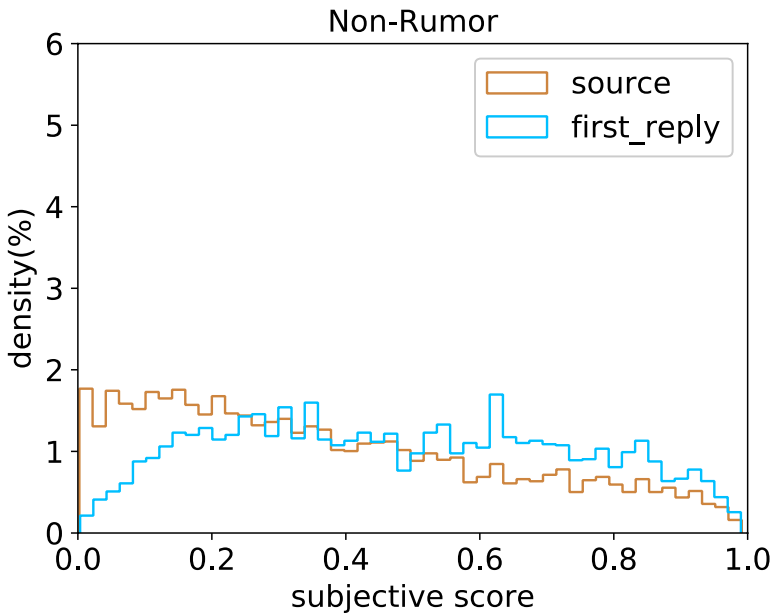
- 主观性信息的偏移在谣言和非谣言的传播过程中的差别



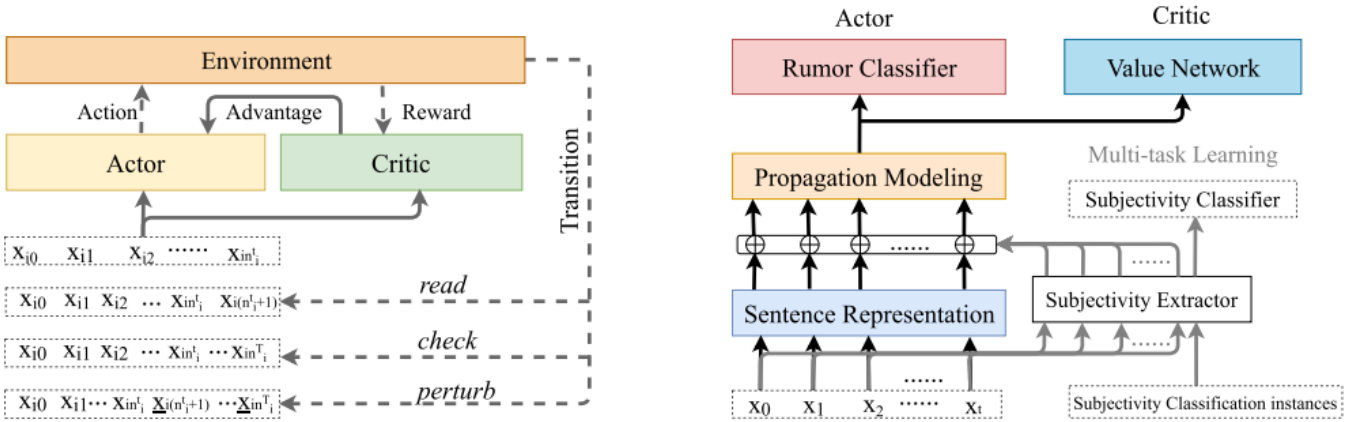
研究动机

- 源消息相比于第一条回复的主观性偏移，说明主观性信息也可以帮助早期检测
- 源信息相比于所有回复的主观性偏移消息





模型



(a) Th entire framework of the SIFTER

(b) The architecture of the actor and the critic in the SIFTER

Fig. 3. The overview of the SIFTER framework. The SIFTER is a sequential training framework implemented by reinforcement learning and consists of an environment, an actor, and a critic. The interaction among the environment, the actor, and the critic is shown on the left figure, and the architecture of the actor and the critic is shown on the right figure.

实验结果

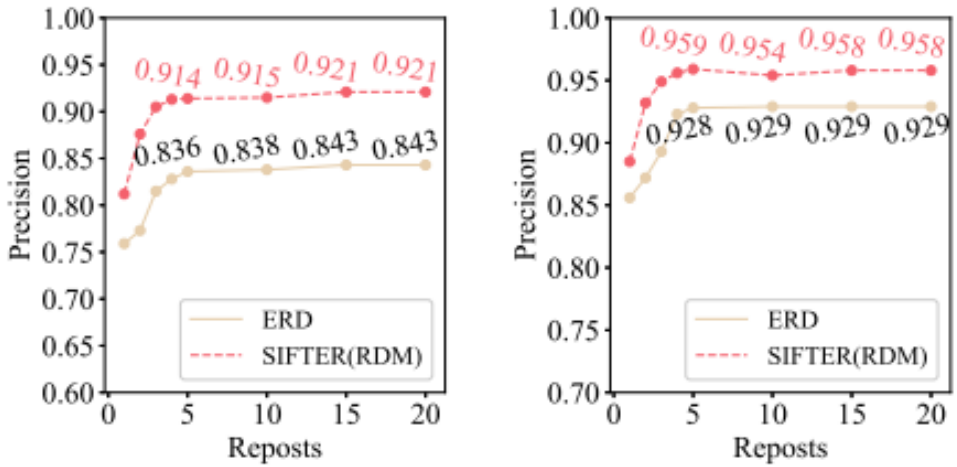
- 谣言检测准确率结果

Models	TWITTER				WEIBO			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
RNN	0.748	0.707	0.659	0.682	0.871	0.816	0.954	0.879
LSTM	0.795	0.719	0.683	0.701	0.907	0.846	0.958	0.899
GRU-2	0.808	0.741	0.694	0.717	0.914	0.876	0.956	0.914
RDM	0.873	0.817	0.823	0.820	0.957	0.950	**0.963**	0.957
ERD	0.858	0.843	0.811	0.826	0.933	0.929	0.936	0.932
SubRRD	**0.894**	**0.921**	**0.933**	**0.927**	**0.971**	0.958	0.962	**0.960**

Table 3: Test Accuracy on TWITTER and WEIBO Data

实验结果

- 谣言早期检测结果



(a) Twitter (b) Weibo

Fig. 4. The precision of the early rumor detection. The y-axis represents the precision score. The x-axis is the maximum number of repost comments in the conversation, indicating how early the model to perform the prediction.

实验结果

- 跨领域谣言检测准确率结果

Model	Charlie.	Ferguson.	German.	Ottawa.	Sydney.	Average
RNN	71.99%	71.27%	55.13%	52.36%	58.27%	61.78%
LSTM	76.63%	69.52%	61.70%	58.20%	62.36%	65.68%
GRU-2	75.19%	70.22%	58.72%	54.83%	60.23%	63.84%
RDM	78.35%	72.41%	61.88%	71.73%	70.64%	71.00%
ERD	76.66%	71.29%	60.62%	70.70%	68.54%	69.56%
SubRRD	**82.94%**	**79.54%**	**71.21%**	**75.36%**	**75.72%**	**76.95%**

Table 4: Cross Domain Accuracy on TWITTER dataset

Model	Politics	Health	Entertain	Average
RNN	75.00%	81.84%	76.36%	77.73%
LSTM	80.71%	85.32%	79.39%	81.81%
GRU-2	80.71%	85.57%	77.57%	81.28%
RDM	86.42%	88.81%	85.45%	86.89%
ERD	84.67%	87.92%	86.66%	86.72%
SubRRD	**91.91%**	**92.38%**	**91.92%**	**92.07%**

Table 5: Cross Domain Accuracy on WEIBO dataset

实验结果

- 谣言检测判别一致性实验结果

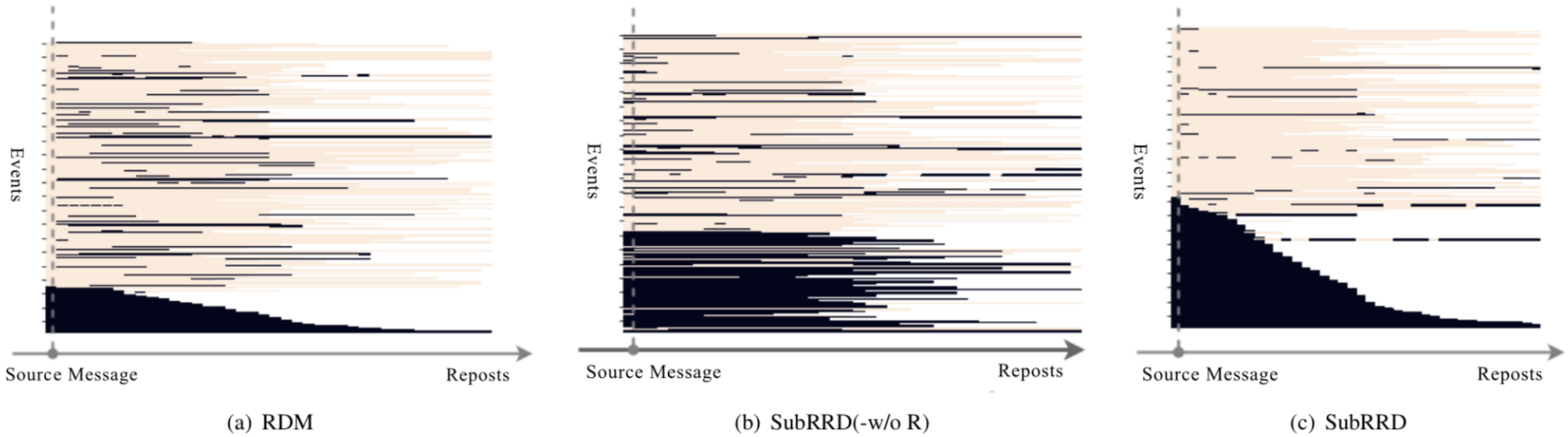


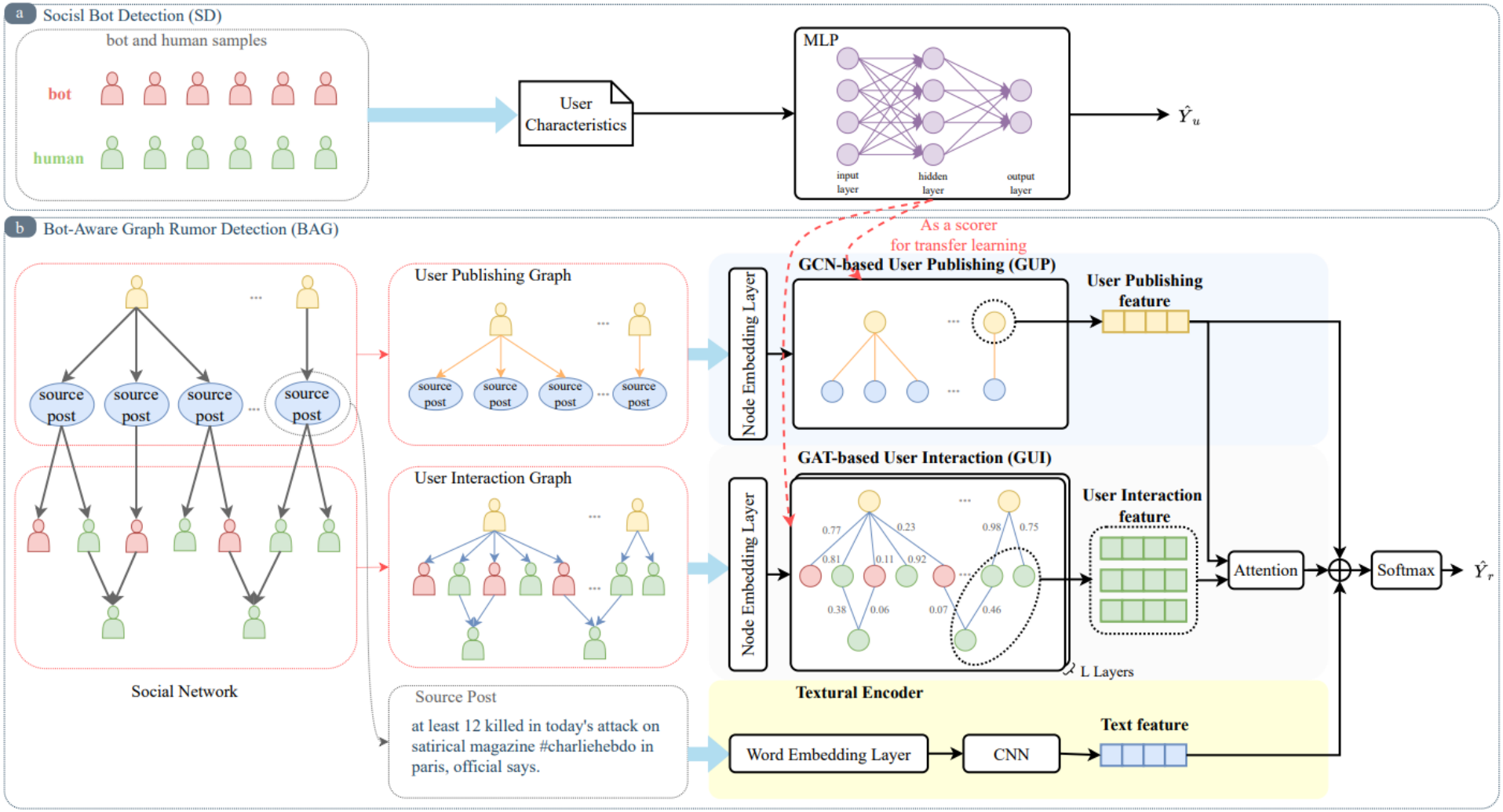
Figure 9: Alarming signal heatmap on TWITTER. In the figure, each row represents an event sequence, where the policy network outputs its decision at each position. If its decision is to trigger a rumor alarm, the corresponding signal on the heatmap will be marked as light color; Otherwise, the corresponding signal will be marked as dark color. Therefore, the inconsistency of color in each row indicates *prediction inconsistency* of the models that its prediction of an event relies on the choice of the checkpoint.

Social Bot-Aware Graph Neural Network for Early Rumor Detection

- 社会学研究表明，谣言在社交媒体传播过程中，存在大量的社交机器人行为，这尤其体现在谣言传播的早期阶段，并对接收源信息的用户产生影响。
  - Shao, 2018. The spread of low credibility content by social bots. Nature communications, 9(1): 1–9.

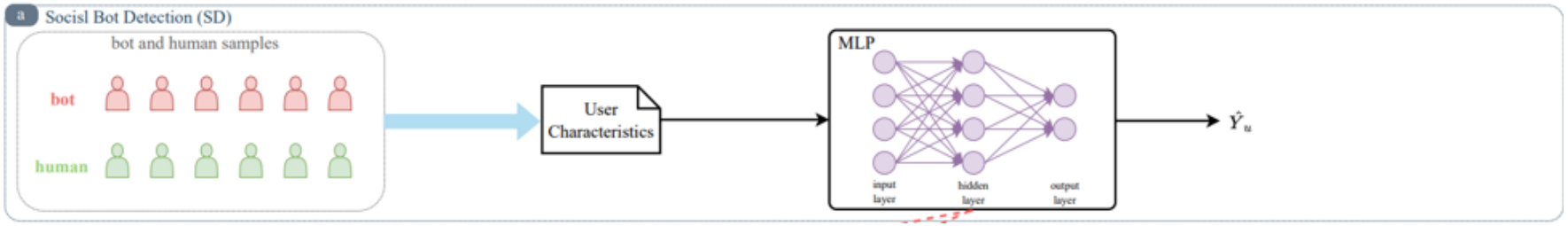
- 因此，对社交机器人进行有效识别，剔除其在谣言检测中的干扰，不仅能够辅助识别谣言，还能提高谣言检测的时效性。

Social bot aware 谣言判别模型



Social bot aware 谣言判别模型

- 在社交机器人判别模块中，重点目标是判断其是否是一个社交机器人及其对于谣言检测的影响，因此主要考虑了用户名长度、粉丝数量等特征。



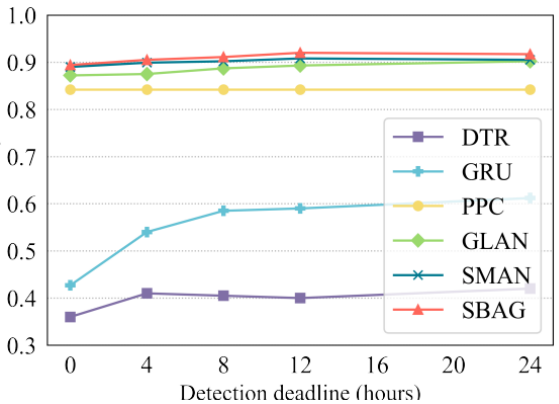
实验结果

- 谣言检测准确率结果

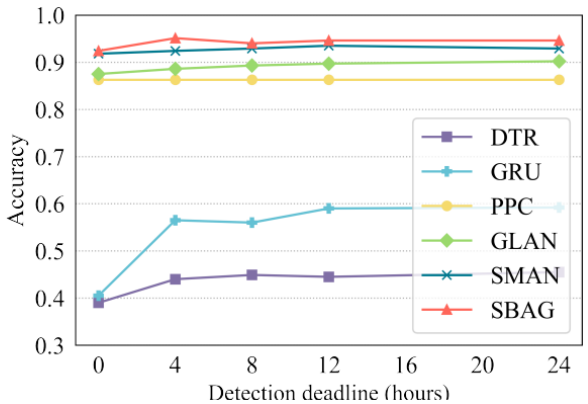
Method	Acc.	NR			FR		
		Precision	Recall	F1	Precision	Recall	F1
DTR	0.732	0.726	0.749	0.737	0.738	0.715	0.726
DTC	0.831	0.815	0.847	0.830	0.847	0.815	0.831
RFC	0.849	0.947	0.739	0.830	0.786	0.959	0.864
SVM-RBF	0.818	0.815	0.824	0.819	0.822	0.812	0.817
SVM-TS	0.857	0.878	0.830	0.857	0.839	0.885	0.861
GRU	0.910	0.952	0.864	0.906	0.876	0.956	0.914
PPC	0.921	0.949	0.889	0.918	0.896	0.962	0.923
GLAN	0.946	0.949	0.943	0.946	0.943	0.948	0.945
SMAN	0.951	0.937	**0.967**	0.952	**0.967**	0.936	0.951
SBAG	**0.957**	**0.967**	0.947	**0.957**	0.947	**0.967**	**0.957**

实验结果

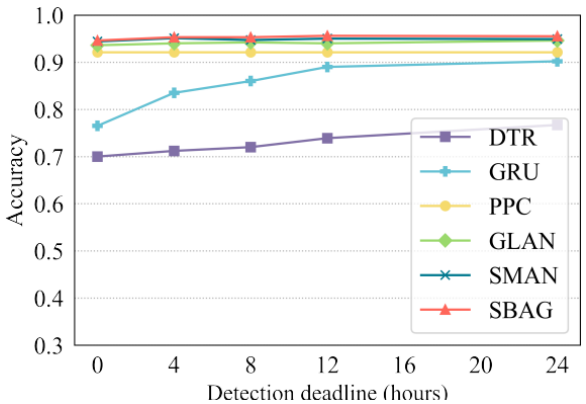
- 谣言检测时效性结果



(a) Twitter15.



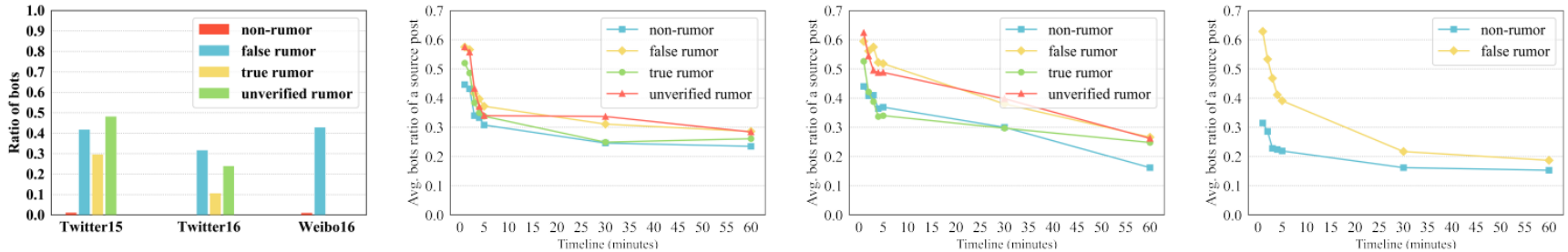
(b) Twitter16.



(c) Weibo16.

实验结果

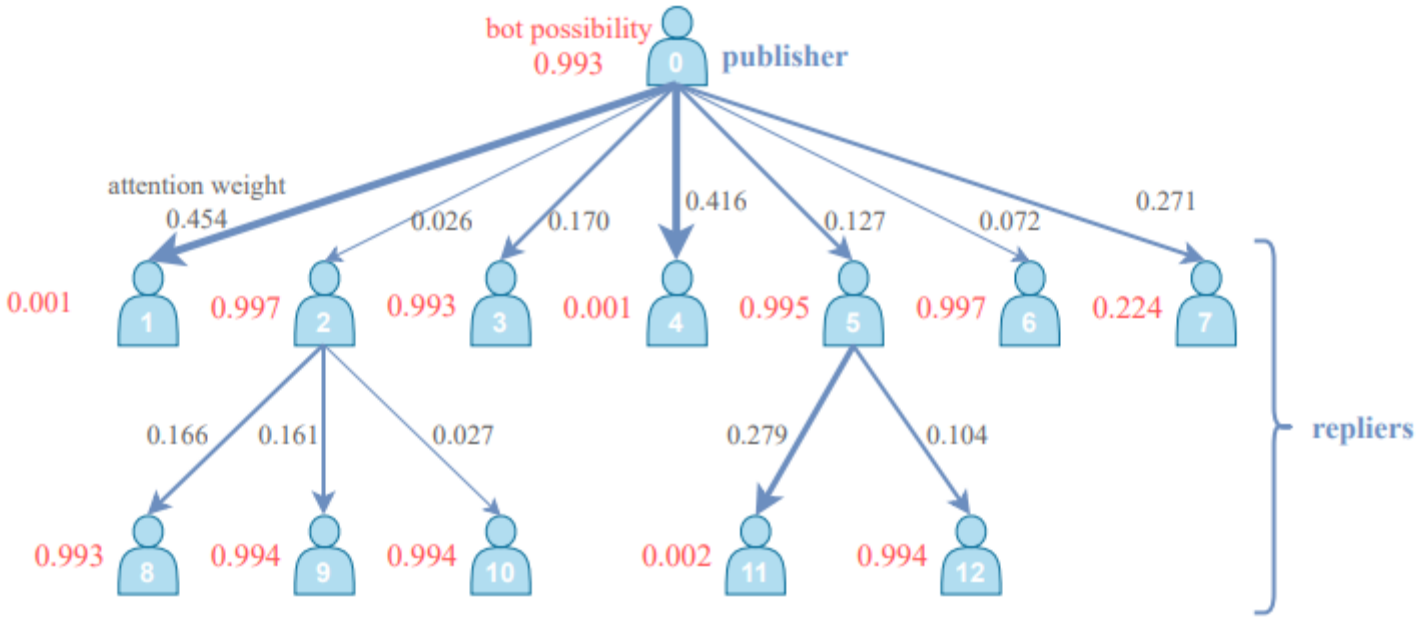
- 谣言检测中社交机器人的影响实验



(a) Relationship between rumors and publishers. (b) Twitter15: Avg. bots ratio per source post. (c) Twitter16: Avg. bots ratio per source post. (d) Weibo16: Avg. bots ratio per source post.

Figure 4: Relationship between rumors and users.

案例分析


bot possibility
0.993
0
publisher
attention weight
0.454
0.026
0.170
0.416
0.127
0.072
0.271
0.001
1
0.997
2
0.993
3
0.001
4
0.995
5
0.997
6
0.224
7
0.166
0.161
0.027
0.279
0.104
repliers
0.993
8
0.994
9
0.994
10
0.002
11
0.994
12

QA?